Language Technologies for All (LT4All). Enabling Language Diversity & Multilingualism Worldwide

Joseph J Mariani, Gilles Adda, Khalid Choukri, Irmgarda Kasinskaite Buddeberg, Hélène Mazo, Sakriani Sakti

LT4All 2019

Language Technologies for All (LT4All)
Enabling Language Diversity & Multilingualism Worldwide

Proceedings

December 4-6, 2019

UNESCO Headquarters, Paris, France

LT4All Sponsors and Organizers

Organized and supported by

in partnership with

Founding public donor

Other public donors

Founding private sponsor

Gold

Silver

Bronze

Travel Grants

Supporters

Endorsers





Order copies of these proceedings from:

European Language Resources Association (ELRA)
9 rue des Cordelières
75013 Paris
FRANCE
Tel: +33 143133333
Fax: +33 143133330
info@elda.org

ISBN: 979-10-95546-33-7
EAN: 9791095546337

"Everyone should have the possibility to get access to Language Technologies in their native languages, including indigenous languages"

Digital technologies, in particular Language Technology, content development and dissemination, play a growing role in influencing societal development and contributing to the intergenerational transmission of Indigenous languages in today's world.

In this context, policy and decision makers, language technology developers, media and information providers, and other relevant public and private stakeholders should be alert and sensitive to barriers which impede the availability of new technology, content and services to Indigenous language users. Provisions should take account of consent considerations and should, where possible, encourage the application of solutions whose delivery is based on open standards – in particular advanced technologies such as Artificial Intelligence.

In this document, Language Technologies (LTs) range from simple keyboarding technologies, spelling/grammar checkers up to speech and speaker recognition, machine translation for text and audio, speech synthesis, spoken dialog, text and document understanding, generation and summarization, sentiment and opinion analysis, answers to questions, information retrieval and knowledge access, sign language processing, etc.

Language Resources cover all types of data sets including audio/video recordings, with or without human annotations, textual corpora, lexica in a machine readable format. They also cover other modalities such as sign languages.

We assume that (1) the needs of each person should be fulfilled and (2) particular attention should be paid to people with specific needs.

The first 10 statements from the Strategic Document Drafting Committee are:

1. In order to break the digital divide, all communities, particularly indigenous communities, should have the possibility to get access to all Language Technologies (LTs) by easily accessing - or producing - multilingual knowledge, communication and education material, and services, in or using their native languages.

2. Native languages should be usable for communication between humans speaking different languages. Cross-lingual LT solutions exist and should be extended to cover all languages, as needed. This will improve mutual understanding while facilitating the access to foreign cultures.

3. Today only 2% of the 7,000+ languages spoken in the world are LT-enabled. Consequently, languages that are not LT-enabled are at risk of never getting access to the digital world and those languages with some digital presence yet weakly LT-enabled may quickly face digital extinction to the benefit of the few so-called "major" languages.

4. Communities should have the capability to define their own needs, expectations and requirements. Today, 20% of the human languages spoken on the planet cover about 99% of the population [1]. It is very likely that LT development efforts depend on the market and community size. It is essential to join forces with experienced communities to pay particular attention to the remaining 80% of the languages[2], irrespective of any consideration on economic interest or community demographics.

[1] https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers
[2] https://www.ethnologue.com/statistics/size

5. In many geographical areas, internet access and digital literacy are strong barriers which have to be seriously addressed, in agreement with the local communities. Deployment of LT applications and services is mainly hindered by the lack of language policies that should foster both collections of needed and sharable language resources and expertise. For instance, having a writing system, which is often the first building block of many technologies, is not always a pre-requisite and can be overcome by resorting to alternative approaches based on spoken information.

6. Most of the key state of the art LTs exist today as open source packages that could be customized and tuned for all languages, assuming that needed language resources and expertise are available. Such trend should be supported and encouraged. Nevertheless, more research is required and so are technology improvement, easy adaptation and portability methodologies, which should be part of a shared research agenda for all languages. The research community should pay attention to languages with no digitized data to undergird participation and decisions about where to invest resources should carefully consider this perspective.

7. Assessment of LT performance for a given language should be systematically conducted to ensure its usability for actual services. LTs that are not mature enough and hence not usable may lead to more confusion than to building actual and valuable services. It is important to ensure that a clear information is conveyed for experimental technologies and prototypes that could be released to the public in languages where performance is not yet demonstrated to reach industry benchmark.

8. Education is a human right. The availability of online courses (e.g. MOOC) on important domains is a major trend that has developed recently but for a limited number of languages. LTs can help develop and improve the access to these online learning programmes for a non-native audience. Such access can build on all modalities such as audio, video, text, sign languages, etc. By extracting resources from archives of texts and stories, from oral or written culture, LTs can also support the development of educational resources in the native languages.

9. Language preservation is part of the culture preservation and human heritage. LTs allow the curation of resources such as audio/video recordings, textual corpora, grammars, and lexica, among others, that are representative of the culture and language use within the community. Through the curation of resources, language documentation activities can be conducted to the benefit of future generations as part of their cultural heritage. Documenting a language also requires to analyse the content collected; LTs can accelerate this process, allowing more resources and more languages to be processed.

10. The revitalization of languages, particularly endangered languages, can be promoted by granting indigenous communities a better access to native resources and knowledge. LTs can also directly help revitalize languages: we can think of using speech synthesis and computer-assisted language learning to allow people in the communities to remain exposed to both the language sounds and to an accurate pronunciation of the language.

Introduction by the Organizers of the Thematic Tracks Achievements and Challenges (Day 2 and Day 3)

Gilles Adda, Khalid Choukri, Irmgarda Kasinskaite, Joseph Mariani, Hélène Mazo, Sakriani Sakti

Information and knowledge in different languages is key to the achievement of sustainable development. The production of new knowledge and adequate responses to emerging challenges, such as poverty, climate change, the digital divide, uneven economic opportunities and social exclusion, will require the preservation of knowledge accumulated in local traditions and practices through the appropriate tools.

In today's world, digital literacy, access to broadband connectivity and quality content, including in local languages, are also prerequisites for the fulfillment of our human rights and fundamental freedoms, as well as for our participation in the development of societies. The denial of such rights and freedoms concerns access to knowledge and information in different languages.

Despite the policies implemented and the technological progress achieved, many language users, particularly Indigenous peoples, still experience barriers to access information online. At school, in the workplace, and in everyday life, they encounter obstacles in using the domain name system, mobile phones, computers, applications and other tools in their own languages. These difficulties are compounded by gender, disability, linguistic, economic, socio-political and digital divides.

In this context, Language Technologies (LT), greatly contribute to the promotion of linguistic diversity and multilingualism in the digital world. These are moving outside research laboratories into numerous applications in many different areas. They include keyboard technologies, spelling/grammar checkers up to speech and speaker recognition, machine translation for text and audio, speech synthesis, and spoken dialogue. They also include text and document understanding, generation and summarization, as well as sentiment and opinion analysis, answers to questions, information retrieval and knowledge access, sign languages processing, etc.

However, a small number of over the 7000 languages spoken around the world have associated Language Technologies, with various levels of quality. The majority of languages can be referred to as under-resourced or as not supported. This situation puts the users of many languages – a vast majority of indigenous languages – in a disadvantageous situation, creating a digital divide, and placing their languages in danger of digital extinction, if not complete extinction.

For over a decade, UNESCO has been committed to promoting a new vision of inclusive knowledge societies, highlighting the importance of creation, dissemination, preservation and utilization of information and knowledge using digital technologies. This approach acknowledges the increasing role of information and knowledge in society, as powerful resources for the creation of wealth, social transformation and human development.

Therefore, UNESCO considers that it is mandatory to support linguistic diversity and multilingualism as essential pillars of fostering pluralistic, equitable, open and inclusive knowledge societies. The Organization also recognizes linguistic diversity and multilingualism as a source of enrichment for humanity and development. It also encourages its Member States to develop comprehensive language-related policies, to allocate resources and use appropriate tools in order to promote and facilitate linguistic diversity and multilingualism, including the Internet and media. In this regard, the Organization through the Recommendation concerning the Promotion and Use of Multilingualism and Universal Access to Cyberspace, promotes the development of multilingual content and

systems, facilitation of access to networks and systems, development of public domain content and reaffirmation of equitable balance between the interests of rights holders and the public interest.

Furthermore, UNESCO as a lead United Nations agency for the organization of the 2019 International Year of Indigenous Languages has worked to raise awareness on the critical status of many indigenous languages around the world, and to mobilize necessary resources for an immediate action as well as the implementation of the Action Plan for organizing the 2019 International Year of Indigenous Languages.

Based on the Roadmap towards UNESCO's World Atlas of Languages, the Organization is already developing the online platform that will contain data on languages, policies, regulations, technical recommendations and best practices in this field. It is expected that a new globally accessible and open online platform will be used for monitoring and promotion of the world's languages online, strengthen cooperation and knowledge sharing. This will be by using open and inclusive technological solutions among international, regional and national language institutions, language users and other public and private partners.

Key messages

The development of Language Technologies provides opportunities to improve the free flow of ideas by word and image in different languages and should leave no one behind regardless of the users' age, gender, abilities, language or location.

All language users should have access to Language Technologies and be able to use them to provide and receive appropriate content and services in their own language.

Equipping language users with the necessary tools to benefit from the latest digital developments requires the joint and long-term efforts of all stakeholders, including governments, language users, in particular Indigenous peoples, academia, civil society and the private sector.

Major objectives, expected outcomes and outputs

Within the framework of the 2019 International Year of Indigenous Languages, the 3-day International Conference looks at the evolving environment of linguistic diversity and multilingualism which plays an important role in our societies. The event also aims to identify recommendations on how to harness technology for the preservation, support and promotion of languages, including lesser-used and indigenous languages, as well as on how to increase and facilitate communication between language users. The Conference will conclude its work with an Outcome Document in order to set out the direction for future global actions in the area of Language Technologies.

With a view to provide access to information and knowledge to all language users and facilitate their inclusion and participation in building sustainable knowledge societies, UNESCO, in close cooperation with the Government of the Khanty-Mansiysk Autonomous Okrug-Ugra (Russian Federation), the European Language Resources Association (ELRA) and its Special Interest Group on Under-resourced languages (SIGUL), and in partnership with UNESCO Intergovernmental Information for All Programme (IFAP) and the Interregional Library Cooperation Centre, as well as with support of other public organizations and sponsors, has organized this conference with the title **"Language Technologies for All: Enabling Linguistic Diversity and Multilingualism Worldwide"** at its Headquarters in Paris, France from 4 to 6 December 2019.

The major objective of the event is to promote the human rights and fundamental freedoms of all language users to access and create information and knowledge in language they best understand and to encourage all relevant stakeholders to take concrete measures for the promotion of linguistic diversity and truly multilingual internet and Language Technologies, with special focus on indigenous languages within the context of the international normative instruments and international cooperation mechanisms such as the 2019 International Year of Indigenous Languages (IYIL2019).

The specific objectives are to:

• Identify existing challenges and barriers for language users to access and create information and knowledge in different languages,

• Explore the relationship between technologies and languages from a scientific, technical, cultural, linguistic, economic and political perspective,

• Look at the socio-cultural impact of technologies on languages and their users, with special focus on lesser-used, minority and indigenous languages,

• Raise awareness of the significance and complexity of linguistic diversity and multilingualism, and the sense of urgency for taking concrete measures and actions by all relevant stakeholders,

• Share good practices on existing technological solutions to tackle the growing linguistic divide between users of different languages.

The expected outputs:

• Concrete recommendations to improve the status of linguistic diversity and multilingualism with regards to the role of language technologies in increasing access to multilingual information and knowledge,

• Research and identification of good practices and technical solutions for the promotion of linguistic diversity and multilingualism

• More public-private stakeholders engaged in the production of multilingual content, the establishment of enabling environments and further deployment of appropriate Language Technologies for the promotion of linguistic diversity and multilingualism,

The expected outcome:

• Long-lasting commitment of all stakeholders to the development of new Language Technologies, as an effective and necessary means for promoting linguistic diversity and multilingualism.

New technology such as Artificial intelligence is used in systems that provide translations between languages, documentation of languages and learning languages through natural language and voice interface platforms. While technology is advancing in artificial intelligence, some languages are disadvantaged by the economic models prevalent in such emerging technologies.

The International conference has therefore also taken stock of the technological state-of-the-art in language technologies, including Artificial Intelligence methods such as Machine Learning, including Deep Learning, and its need for linguistic data.

The 3-day conference was composed of keynote talks, oral and poster presentations, panel discussions as well as reports, demonstration of systems and innovative applications, and experiences gained in the deployment of technologies in promoting linguistic diversity and multilingualism. Innovative solutions were presented at the exhibition specially installed during the event.

The event will draw conclusions and recommendations in an **Outcome Document** resulting from the presentations, panel discussions and written contributions.

The conference gathered UNESCO Member States, international bodies, regional governments and administrations, academia, language technology researchers, linguists, industry, indigenous peoples' and language policy and decision makers from around the world.

The participation to the conference was on invitation only. The conference was well attended by 400+ participants.

Committees

The International Conference was organized by UNESCO, Government of the Khanty-Mansi Autonomous Area – Ugra and the European Language Resources Association (ELRA) and its Special Interest Group on Under-resourced languages (SIGUL), in partnership with the UNESCO Information for All Programme (IFAP) and Interregional Library Cooperation Centre.

The Conference was conducted by a **Coordination Committee** that is chaired by the Assistant Director-General for Communication and Information of UNESCO. Other members of the Organizing Committee include a representative from the:

- Government of the Khanty-Mansi Autonomous Area – Ugra (Mr Alexey Shipilov, First Deputy Governor of the Khanty-Mansiysk Autonomous Area – Ugra),
- European Language Resources Association – (Mr Khalid Choukri, General Secretary, European Language Resources Association ELRA, Mr Joseph Mariani, Honorary President of ELRA, and Ms Sakriani Sakti, Secretary, Joint Special Interest Group on Under-resourced languages, SIGUL),
- UNESCO Intergovernmental Information for All Programme (Ms Dorothy Gordon, Chair and Mr Evgeny Kuzmin, Vice-Chair),
- Steering Committee for the organization of the 2019 IYIL2019 (Ms Aili Keskitalo, Indigenous Co-chair and Chair of the Ad-hoc group on Language Technologies),
- UNESCO Secretariat (Ms Irmgarda Kasinskaite-Buddeberg and Mr Jaco du Toit).

The Coordination Committee is supported by sub-committees for Day 2 and Day 3:

- International Honorary Advisory Committee,
- Program Committee,
- Organizing Committees,
- Outcome Document Drafting Committee.

International Honorary Advisory Committee for the thematic tracks Achievements and Challenges:

Ahmed Boukous, Rector, Institut Royal de la Culture Amazighe (IRCAM), Morocco
António Branco, Professor, University of Lisbon & President of the European Language Resources Association (ELRA), Portugal
Nicoletta Calzolari, Research Fellow, Institute for Computational Linguistics "A. Zampolli - CNR & Honorary President of the European Language Resources Association (ELRA) and LREC General Chair, Italy
Myrna Cunnningham, President of the Fund for development of Indigenous Peoples of Latin America and the Caribbean (FILAC) & Former Chair, UN Permanent Forum on Indigenous Issue, Nicaragua
Lang Fafa Dampha, Executive Secretary, African Academy of Languages (ACALAN), African Union Commission, Mali
Attie de Lange, Director, South African Centre for Digital Language Resources (SADiLaR), South Africa
Véronique Delvaux, FNRS Research Associate, University of Mons & President, Association Francophone de la Communication Parlée (AFCP), Belgium
Jill Evans, European MP for Wales, UK
Vigdís Finnbogadóttir, Former President of the Republic of Iceland & UNESCO Goodwill Ambassador for Language Diversity, the Vigdís Finnbogadóttir Institute of foreign Languages, University of Iceland, Iceland
Martine Garnier-Rizet, Head of Digital Technology and Mathematics Department, ANR, France
Dorothy Gordon,Chair, Intergovernmental Council, UNESCO Information for All Programme (IFAP), Ghana
John Hansen, Professor, University of Texas at Dallas & President, International Speech Communication Association (ISCA), USA
Aili Keskitalo, President, Sámi Parliament, Norway
Mark Liberman, Professor, University of Pennsylvania & Director, Linguistic Data Consortium (LDC), USA
Joseph Lo Bianco, Former President, Australian Academy of the Humanities & Professor of Language and Literacy Education, University of Melbourne, Australia
Sixto Molina, Head of Secretariat, European Charter for Regional or Minority Languages DGII & Directorate of Anti-Discrimination, Council of Europe, France
Makoto Nagao, Professor Emeritus, Kyoto University, Member of the Japan Academy, Japan
Satoshi Nakamura, Professor, Nara Institute of Science and Technology (NAIST) & General Convenor, Cocosda & Oriental Cocosda, Japan
Dina Ocampo, Former undersecretary, DepED (Department of Education) & Professor, College of Education, University of the Philippines, the Philippines
Patrick Paroubek, Senior Research Engineer, LIMSI-CNRS & President, Association pour le Traitement Automatique des Langues (ATALA), France
Adama Samassekou, Founding Executive Secretary, AU/ACALAN, President, WSIS PrepCom for the Geneva Phase, President, MAAYA the World Network for Linguistic Diversity & Former Minister of Education, Mali
Algirdas Saugardas, Former European Deputy & former Minister of Foreign Affairs, Lithuania
Paul de Sinety, Délégué Général à la Langue Française et aux Langues de France, DGLFLF, France
Ming Zhou, Principal researcher, Microsoft Research Asia & President, Association of Computational Linguistics (ACL), China

Chengqing Zong, Research Fellow, Chinese Academy of Sciences & President, Asian Federation of Natural Language Processing (AFNLP), China

Program Committee for the thematic tracks on the Achievements and Challenges:

Tunde Adegbola, Executive Director, African Languages Technology Initiative (Alt-i), Nigeria
Mirna Adriani, Dean, Computer Science Faculty, Universitas Indonesia, Indonesia
Shyam Agrawal, KIIT Group of Colleges, Gurgaon, India
Ahmed Ali, Qatar Computing Research Institute, Hamad Bin Khalifa University, Qatar
Fadoua Ataa Allah, Institut Royal de la Culture Amazighe (IRCAM), Morocco
Antti Arppe, Alberta Language Technology Laboratory (ALTLab), University of Alberta, Canada
Kalika Bali, Microsoft Research, India
Nicholas Barla, Indigenous Peoples Forum Odisha (IPFO), India
Dorothee Beermann, Norwegian University of Science Technology (NTNU), Norway, Polytext AS, Norway
Martin Benjamin, Kamusi Project International, Switzerland
Laurent Besacier, Laboratoire d'Informatique de Grenoble (LIG), France
Brigitte Bigi, Laboratoire Parole et Langage (LPL), France
Steven Bird, Charles Darwin University, Australia
Sonia Bosch, University of South Africa, South Africa
Karim Bouzoubaa, Department of computer science, Mohammadia School of Engineers, Mohammed V, University in Rabat, Morocco
Alena Butryna, Google Research, USA
Chris Cieri, Executive Director, Linguistic Data Consortium (LDC), UPenn, USA
Franciska de Jong, Executive Director, CLARIN European Research Infrastructure for Language Resources and Technology, The Netherlands
Emmanuel Dupoux, EHESS, France
Salwa el Ramly, President, Egyptian Society of Language Engineering (ESOLE), Egypt
Vicent Fenollar i Sastre, Policy and Outreach Manager, Network to Promote Linguistic Diversity (NPLD), Belgium
Colleen Fitzgerald, Former Director for Documenting Endangered Languages (DEL-NSF), University of Texas at Arlington, USA
Mikel Forcada, Universitat d'Alacant, Spain
Philippe Gelin, Head of Sector Multilingualism, DG/CONNECT - European Commission, Belgium
Paul Geraghty, University of the South Pacific, Suva, Fiji
Dafydd Gibbon, Bielefeld University, Germany & Jinan University Guangzhou, China
Lea Gimpel, Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH, Germany
Thibault Grouas, Délégation Générale à la Langue Française et aux Langues de France (DGLFLF), France
Ximena Gutierrez-Vasques, Universidad Nacional Autonoma de Mexico, Mexico
Dieter Halwachs, Plurilingualism Research Unit, University of Graz, Austria
Thomas Hanke, University of Hamburg, Germany
Mary Harper, USA
Mark Hasegawa-Johnson, University of Illinois, USA
Auður Hauksdóttir, The Vigdís Finnbogadóttir Institute of Foreign Languages, University of Iceland, Iceland
Julia Hirschberg, Columbia University NYC, USA
Misako Ito, Regional Adviser for Communication and Information, UNESCO Bangkok,

Thailand
Mustafa Jarrar, Computer Science Dept Birzeit University, Palestine
Kristiina Jokinen, University of Helsinki, Finland
Alexey Karpov, St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), Russia
Sanjeev Khudanpur, Johns Hopkins University, USA
Sabine Kirchmeier, Vice-President, EFNIL (European federation of National Institutions for Languages), and Danish Language Council, Denmark
Kate Knill, Machine Intelligence Laboratory, Cambridge University, UK
Andras Kornai, Computer Science Research Institute, Hungarian Academy of Sciences (MTA SZTAKI), Hungary
Simon Krek, Jožef Stefan Institute, Slovenia
Terry Langendoen, University of Arizona, USA
Eirik Larsen, Political Advisor, Sámi Parliament, Norway
Gina-Anne Levow, University of Washington, USA
Haizhou Li, National University of Singapore (NUS), Singapore
Aijun Li, Institute of Linguistics Chinese Academy of Social Sciences, China
Lydia Liu, Institute for Comparative Literature and Society (ICLS), Columbia University, USA
Sjur Nørstebø Moshagen, Head of Sámi language technology application development, UiT The Arctic University of Norway, Norway
Shri Narayanan, University of Southern California, USA
Girish Nath Jha, Jawaharlal Nehru University (JNU), India
Graham Neubig, Carnegie Mellon University, USA
Ailbhe Ní Chasaide, Trinity College Dublin, Ireland
Nathaniel Oco, the Philippines
Nicholas Ostler, Chair, Foundation for Endangered Languages (FEL), UK
Win Pa Pa, University of Computer Studies Yangon (UCSY), Myanmar
Cecilia Piaggio, Founder, LatinoAmericaHabla, Argentina
Thierry Poibeau, Ecole Normale Supérieure Paris, France
Delyth Prys, Bangor University, Wales, UK
Georg Rehm, DFKI Berlin, Germany
Michael Rießler, University of Eastern Finland, Finland
Clara Rivera, Google Research, UK
Mike Rosner, Department of Artificial Intelligence, University of Malta, Malta
Priyankoo Sarmah, Indian Institute of Technology Guwahati, India
Yusuf Sawaki, Center for Endangered Languages Documentation (CELD), Universitas Negeri Papua, Indonesia
Odette Scharenborg, Delft University of Technology, The Netherlands
Holger Schwenk, Facebook AI Research, France
Mandana Seyfeddinipur, SOAS Endangered Languages Documentation Programme, University of London, UK
Tanja Schultz, Cognitive Systems Lab (CSL), Universität Bremen, Germany
Claudia Soria, Istituto di Linguistica Computazionale "A. Zampolli"- CNR, Italy
Virach Sornlertlamvanich, SIIT Thammasat University, Thailand & Musashino University, Japan
Juan Steyn, South African Centre for Digital Language Resources (SADiLaR), South Africa
Sebastian Stüker, Karlsruhe Institute of Technology, Germany
Akatsuki Takahashi, Advisor for Culture at UNESCO Office for the Pacific States, Samoa
Nick Thieberger, The Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC), University of Melbourne The ARC Centre of Excellence for the Dynamics of

Language, Australia
Charu Bikash Tripura, Regional Capacity Building Coordinator Asia Indigenous Peoples Pact (AIPP), Thailand
Jan Trmal, Center for Language and Speech Processing (CLSP), Johns Hopkins University, USA
Alexey V. Tsykarev, Chair, Center of Support of Indigenous People and Civic Diplomacy "Young Karelia", Former Chair, UN Expert Mechanism on the Rights of Indigenous People & Incoming Expert Member, UN Permanent Forum on Indigenous Issues, Russia
Francis Tyers, Indiana University, USA
Charl van Heerden, Saigen, South Africa
Zygmunt Vetulani, Adam Mickiewicz University in Poznan, Poland
Lining Wang, Center for the Protection and Research of Language Resources of China Beijing Language and Culture University (BLCU), China
Daniel Willett, Amazon Alexa R&D, Aachen, Germany
Masahiro Yamada, National Institute for Japanese Language and Linguistics (NINJAL), Japan

Organizing Committee for the thematic tracks on the Achievements and Challenges:

Gilles Adda, LIMSI-CNRS, France
Khalid Choukri, ELRA, France
Irmgarda Kasinskaite-Buddeberg, UNESCO, France
Joseph Mariani, LIMSI-CNRS, France
Hélène Mazo, ELRA, France
Sakriani Sakti, NAIST/RIKEN AIP/SIGUL, Japan

Format and structure

The International Conference programme of three days, consisted of plenary oral and poster sessions, introductory presentations and panel discussions in each session, as well as the exhibition on innovative Language Technologies and supporting events (such as the Welcome Reception).
An Outcome Document will capture the main debates and discussions and will seek to draw concrete recommendations and solutions for informed decision-making regarding international cooperation, research and development. The document will be globally distributed among various stakeholders and integrated into new follow-up activities, including to the 2019 International Year of Indigenous languages.

Language Technology exhibition

The exhibition showcased Language Technology solutions by both public and private partners, with relevance to the theme of the conference.

Program Overview:

Day 1. CONTEXT. Multilingualism for building knowledge societies

Opening Session
Keynote address: Adama Samassekou, H.E. Ibrahim Albalawi
Moderator: Evgeny Kuzmin

Oral Session 1: The multilingual landscape and Indigenous Languages
Moderator: Gilvan Muller de Oliveira

Oral Session 2: Languages and Communication in the 21st century
Moderator: Alisher Ikramov

Oral Session 3: Language policies and globalization: challenges and opportunities
Moderator: Anuradha Kanniganti

Oral Session 4: Multilingualism in education and research
Moderator: Coetzee Bester

Welcome Reception

Day 2. ACHIEVEMENTS. Applying Language Technologies for linguistic diversity and multilingualism

Opening Session
Keynote: Daan van Esch
Moderators: Marko Grobelnik, Heather Souther.
Rapporteur: Eugenia Urrere

Poster Session 1: Pacific Languages
Moderators: Steven Bird, Apolonia Tamata, Isabella Shields

Oral Session 5: Innovative aspects related to applications of Language Technologies in various areas, products and services
Moderators: Febe de Wet, Hermann Ney.
Rapporteur: Ethel Ong

Poster Session 2: European and Arctic Languages
Moderators: Teresa Lynn, Daniil Kocharov.

Oral Session 6: Scientific aspects related to the state of the art in Language Technologies, for spoken, written and sign languages
Moderators: Volker Steinbiss, Ximena Guttierez-Vasques.
Rapporteur: Dessi Puji Lestari.

Poster Session 3: Latin America and the Caribbean Languages
Moderators: Francisco Cláudio Sampaio de Menezes, Marco Antonio Martínez Pérez, Anuschka van 't Hooft

Oral Session 7: Infrastructural aspects
Moderators: Mark Liberman, Amanda Harris.
Rapporteur: Alexey Karpov

Social event
Invited Talks Tunde Adegbola,
Rory O'Connor
Moderators: Sakriani Sakti, Khalid Choukri, Joseph Mariani.

Day 3. CHALLENGES. Addressing the digital divide and multilingualism

Keynote: Lorna Williams
Moderators: Justus Roux, Sanjeev Khudanpur.
Rapporteur: Priyankoo Sarmah.

Oral Session 8: Minority and Indigenous Languages
Moderators: Yoshinori Sagisaka, Sonja Bosch.
Rapporteur: Claudia Soria

Poster Session 4: African Languages
Moderators: Dorothy Beermann, Joseph Nkonga, Audace Niyonkuru

Oral Session 9: Activities for language preservation, reclamation, and enhancement
Moderators: Cyntia Montaño, Mikel Lorenzo Forcada Zubizarreta.
Rapporteur: Satoshi Tamura

Poster Session 5: Asian Languages
Moderators: Budi Irmawati, Netra Mani Rai

Oral Session 10: Scientific aspects related to handling language diversity
Moderators: Laurent Besacier, Gilles Adda.
Rapporteur: Nathaniel Oco

Poster Session 6: USA and Canada Languages
Moderators: Chris Cieri, Francis Tyers, Jan Trmal

Oral Session 11: Developing Language Technologies for All : Best Practices
Moderator: Delyth Prys.
Rapporteur: Emiliana Cruz

Oral Session 12: The Future of Language Technologies for All: Outcome Document and Recommendations and Closing Remarks

Opening speech from Dr Moez Chakchouk, Assistant Director-General for Communication and Information of UNESCO (December 5, 2019)

Good morning,

- Excellencies
- Distinguished guests
- High Representatives of China, France, New Zealand, Norway
- Ladies and gentlemen

It is my great pleasure to address you again, on the second day of this International Conference. Thanks to all of you for being here today, despite current transportation challenges.

I would also like to repeat our gratitude to:

- the Government of the Khanty-Mansiysk Autonomous Okrug-Ugra (Russian Federation);
- the European Language Resources Association (ELRA) and its Special Interest Group on Under-resourced languages (SIGUL);
- UNESCO Intergovernmental Information for All Programme (IFAP);
- the Interregional Library Cooperation Centre;
- as well as additional public and private organizations and our sponsors.

Your invaluable institutional and financial support has enabled this timely event.

This conference is organized in the context of the 2019 International Year of Indigenous Languages, proclaimed by a UNGA resolution that appointed UNESCO as lead UN Agency for its implementation. This international year has aimed to raise awareness about the critical status of indigenous languages worldwide, and the need to mobilize stakeholders and resources to promote, revitalize and support them.

In an increasingly digitalized world, such outcomes can only be reached if concrete efforts are made to close the existing digital gap between dominant, official languages and minority, lesser-used and indigenous ones.

Throughout this international year, nearly nine-hundred (900) events have been taking place all over the world. More than eighty-one thousand (81,000) individuals have joined the global online community of the dedicated website, as well as on social media.

Throughout 2019, UNESCO has carried out a series of regional consultations to identify recommendations for future actions. These have been integrated in the Global Strategic Outcome Document for the International Year. This Document includes a key conclusion on Language Technologies' role in influencing development and contributing to intergenerational transmission of indigenous languages. Now, as the International Year is drawing to a close, this international conference at UNESCO represents a unique occasion to reflect on future actions in the area of Language Technologies.

For our part, since several decades, UNESCO has recognized linguistic diversity and multilingualism as a source of both enrichment for humanity and development. In 2003, our Member States' agreed on the only normative instrument that currently exists at the UN level for the promotion of linguistic diversity. It is called the *Recommendation concerning the Promotion and Use of Multilingualism and Universal Access to Cyberspace.* It is a call for governments to develop comprehensive language-related policies, and to allocate resources and use appropriate tools, including the Internet and media.

There is indeed new potential today, thanks to technological advance. The engagement of major tech and IT companies in the promotion of linguistic diversity and multilingualism is vital for the future of inclusive communication that does not leave indigenous languages behind.

Let us therefore also keep awareness that relatively few Language Technologies have been developed in lesser-used, minority, and indigenous languages, partly due to the unavailability of rich datasets in these languages. Additionally, speakers of these languages do not often have the means and skills to develop strategies to promote their own languages online, or their integration into educational programs and cultural industries.

To understand and map the imbalances, in 2020, UNESCO will launch an online platform for the World Atlas of Languages, as a repository for linguistic diversity and multilingualism. It will be based on a global data collection initiated by UNESCO in close cooperation with the UNESCO Institute of Statistics (UIS), and on the work of leading experts. Next year, we as UNESCO will also publish a World Report on the status of languages, which will include new information and detailed data on human languages, including official and unofficial, spoken as well as sign languages. We hope these will be of use to you going ahead.

Meantime, I am confident that you, the experts and professionals, gathered during these days will devise strategies to address the many challenges to ensure the development and mainstreaming of Language Technologies in minority, lesser-used languages, as well as the main ones. Such multi-stakeholder partnership is one of the key pillars on which the Action Plan of the International Year is based.

This spirit of co-operation should inform new research directions and business strategies to help reduce the cost for developing language technologies, and to encourage an increase in knowledge-sharing practices. This will be the role of the forward-looking Outcome Document of your deliberations. Your recommendations in this Outcome Document will help shape cooperation, research and development.

Your recommendations will also be significant in the light of the proposed declaration of a decade of indigenous languages between 2022 – 2032, which will be decided by the United Nations General Assembly (UNGA) later this month. In other words, the future is there for us to make.

I thank you for your attention and wish you fruitful deliberations.

Moez Chakchouk, Assistant Director-General for Communication and Information of UNESCO

Speech from Dr Khalid Choukri, ELRA Secretary General (December 5, 2019)

Excellencies
Ladies and gentlemen
Dear Colleagues

I am very honored to welcome you all on behalf of the European Language Resources Association. Being a member of an indigenous community myself makes this conference a very special event.

I am very honored to be here today to share with you both our hopes, our plans and our objectives, in organizing and supporting this first **Language Technologies for All** conference. We are confident that we are setting a major sustainable milestone that we will develop for the years to come.

I am responsible for the operations of the European language Resources association and would like to take this opportunity to share with you some of our visions.

The European language resource association (ELRA) is a non-for profit organization based in Europe and acting globally, set up in 1995 (this year we celebrate our 25th anniversary) with the primary mission to support the development of Language Technologies by making language resources available and shareable.

ELRA is a membership-based organization and we are very proud to have among our 50+ members some of the key players from industry and academia but also over 1000 individual members from all areas of Language Technologies, worldwide.

When analyzing the main stream of Language Technologies, we see that today's key paradigm, (and we will review the situation over the next two days), is the data driven paradigm. Whether we talk about Machine Learning, Deep Learning or simply Artificial Intelligence, the main fuel remains data, and we often state that the best data is more data.

ELRA has managed to secure more than 1,400 resources covering. All modalities from Audio/visual, to text, terminology, lexicographical data, sign language resources, data for OCR, and many more modalities, to ensure that several Language Technologies can be developed.

Unfortunately, when I look through the ELRA language resources catalogue but also other data centers, I see that the number of languages we serve is very low; not more than 100 and not all of them with enough data for the development of real applications. This shows the efforts needed to cover even very very partially the needs of the 7000 human languages.

To ensure that the identified resources are usable, we pay serious attention to the ethical as well as the legal aspects when collecting and processing such data.

Our expertise in clearing all Intellectual Property Rights (IPR) helps our partners to have a legal access to the resources they need but also ensures that the right holders get the right credit and acknowledgment, in particular when resources are shared for free. Addressing these IPR issues at the early stages of data collection and production foster the use and reuse of language resources.

In addition to our mission of identifying language resources, negotiating the distribution agreements and licenses, ELRA also produces resources on-demand both for research and commercial applications, very often in partnership with our local colleagues.

To illustrate this production process I would like to highlight the role of public policies and exemplify this through two cases.

Almost 20 years ago, the European Union decided to co-fund the development of speech resources for most of the European Union languages. We established a consortium that collected data for telephony, automotive, and entertainment applications, among others, and extended the language coverage to Latin America, North Africa, and Asia. These resources are very likely an important trigger of some of the today's speech interaction and dictation applications we see on our smartphones.

The second scenario is also a European Commission funded project initiated a few years ago in the framework of a European regulation called "Public sector information directive". Under this directive, all public data have to be made available both for research and industry. An important part of the data that we collected with our partners of the European Language Resource Consortium consists of parallel corpora with aligned multilingual texts that have been used to improve the performance of the Neural-based Machine Translation services used by the European public bodies.

Why am I quoting these cases?

The main reason is to highlight the need for public policies but also for public-private partnerships to produce the required resources. The production processes have gained a tremendous efficiency and cost-effectiveness today compared to 20 years ago, which can be put forward as an argument for investments in Language Resources.

In the first year of ELRA, while working together with another European network, ELSNET, we introduced the concept of Basic LAnguage Resource Kit (BLARK), a Language Resource and Language Technology package that would comprise a minimal set of resources that can encourage researchers to tackle the needs of any language. ELRA advocates for open source policy, in particular for the under-resourced Languages and the indigenous languages.

Another important achievement of ELRA is the set-up of a major conference, the Language Resources and Evaluation Conference (LREC) that focuses on Language Resources and evaluation of technologies. LREC, set up in 1998, takes place every two years. The next edition was due to take place in May 2020, here in France, in Marseille. It will be held in Marseille in 2022.

Now, LREC steadily attracts 1300+ participants from all over the world, who discuss all issues related to language resources I mentioned before, and more. ELRA brings together experts from all Language Technology areas such as speech interaction, spoken dialogue, speech to speech translation, machine translation, sentiment analysis, sign language issues, emotion and affective computing, optical character recognition, text analysis, lexicography terminology, etc.

But in addition to the main conference, LREC makes room to a large number of satellite workshops and tutorials which are smaller events (with 20 to 50 experts getting together) to discuss and explore more specific topics from sign language issues to global lexicography to Arabic, Semitic, Indian, African languages, etc.

During LREC we also cover topics that we will address today and tomorrow with sessions dedicated to infrastructures or strategic roadmaping.

ELRA is also partnering with other organizations to create more synergies. To this end, ELRA and ISCA (international speech Communication association) have established a joint special interest group on Under-Resourced Languages. The core mission of SIGUL, this Special interest Group, is to support linguistic diversity through language technologies and promotes research activities for languages that don't benefit of substantial resources in the Digital world. The special interest group organizes workshops and tutorials to develop its mission. With over 300 members from 34 countries we are looking forward to high-value achievements in the near future.

ELRA feels very proud to see the hundreds of languages that have now access to technology and measure how far we have come. But also foresees the road ahead, a long and difficult one, that will require the involvement of all of us: citizens from indigenous populations, civil society/NGO, as well as government and policy makers. Under a clear ethical statement that indigenous communities should have the capability to define their own needs, expectations, and requirements.

We are very proud to join forces with UNESCO and Khanty-Mansi Region to bring in more experts and language activists to help us draw attention to the requirements and the expectations of indigenous peoples for the languages that don't have all the resources needed by today's technology.

I wish you, I wish us, a very fruitful Language Technologies for All conference.

Khalid Choukri, ELRA Secretary General and ELDA CEO

Table of Contents
Day 2 and Day 3

Session 5. Innovative aspects related to applications of Language Technologies in various areas, products and services

Session 6. Scientific aspects related to the state of the art in Language Technologies, for spoken, written and sign languages

Session 7. Infrastructural aspects

Session Posters

Proceedings, pages 1–4
Paris, UNESCO Headquarters, 5-6 December, 2019.

Speech Technology for Swedish: Current Impact Areas for Applications and Edyson, an Innovative Tool for Accessing Speech Data

David House, Per Fallgren, Jens Edlund
Division of Speech, Music and Hearing, KTH (Royal Institute of Technology)
Lindstedtsvägen 24, 100 44 Stockholm, Sweden
davidh@speech.kth.se, perfall@kth.se, edlund@speech.kth.se

Abstract

This paper presents four of the impact areas for applications in speech technology identified by the National Swedish Language Bank (Språkbanken) and currently being explored by the Speech Section of the Language Bank (Språkbanken Tal). The four areas defined are Cultural Heritage, Inclusion and Accessibility, Health and Aging, and Digitalization. In addition, the paper introduces Edyson, a tool developed at KTH and used for accessing large quantities of speech data. Current use of Edyson in a research project relating to Cultural Heritage and historical speech recordings is presented and discussed.

Keywords: speech annotation, speech browsing, cultural heritage recordings

Résumé

I detta dokument presenteras fyra viktiga områden inom talteknologi som har definierats av Nationella Språkbanken som angelägna och som kartläggs av Språkbanken Tal. De fyra områdena är Kulturarv, Inkludering och Tillgänglighet, Hälsa och Åldrande och Digitalisering. Dessutom introduceras Edyson, ett verktyg utvecklat vid KTH och som används för att få åtkomst till stora mängder taldata. Nuvarande tillämpning av Edyson i ett forskningsprojekt som rör kulturarv och historiska talinspelningar beskrivs och diskuteras.

1. Introduction

The National Swedish Language Bank (Språkbanken) was inaugurated nearly half a century ago, in 1975, at the University of Gothenburg's Department of Swedish. It has since been a nationally and internationally acknowledged research unit with a focus on language resources and language technology. In 2014, Sweden received funding from the Swedish Research Council to join the European language infrastructure Clarin ERIC, which supports language technology in the humanities and the social sciences, and SweClarin was formed. In 2018, the national research infrastructure National Swedish Language Bank (Nationella Språkbanken) was awarded funding from the Swedish Research Council, which also secured the continuation of Swe-Clarin.

In addition to providing for the continued operation of the original Språkbanken (now Språkbanken Text), Nationella Språkbanken adds two new branches: Språkbanken Sam (Eng. "Society") and Språkbanken Tal (Eng. "Speech"). Språkbanken Sam is operated by the Swedish Language Council at the Institute for Language and Folklore (ISOF), which supports research on the languages, dialects and other parts of the intangible cultural heritage in Sweden; and Språkbanken Tal is operated by the Division of Speech, Music and Hearing at KTH, which caters for resources on speech, speech science, and speech technology.

This paper presents four of the impact areas identified as important for the development of applications using speech technology in Swedish: Cultural Heritage, Inclusion and Accessibility, Health and Aging, and Digitalization. Examples of current activities in each area are presented including a more detail description of the tool Edyson, used for accessing large quantities of speech data specifically within the area of Cultural Heritage.

2. Four impact areas

2.1 Cultural Heritage

Cultural Heritage is an important area both on a national and a European scale**:** according to EU Commissioner Carlos Moedas (in charge of Research, Science and Innovation) positions innovation in cultural heritage in "*the intersections between old and new, between physical and digital, and between disciplines*". The European Strategy Forum on Research Infrastructures (ESFRI) has identified CLARIN ERIC (Common Language Resources and Technology Infrastructure) as one of two Research Infrastructures (RIs) of pan-European interest that meet the long-term needs across all scientific areas including social and cultural innovation. Speech technology has the potential to play a key role in cultural heritage, an area where speech is prevalent, but rarely approached with objective and efficient tools.

One of the projects currently running at KTH is a project that develops and makes available speech-to-text, or speech recognition, that is specifically adapted to work on archive materials. There are huge sets of speech and audio data in Swedish archives that cannot be used because they cannot be indexed, simply due to their size. The project produces speech technology research impact in the *Cultural heritage* area by making digitized audio materials truly available digitally, and pushes Swedish speech recognition forward by developing analysis and adaptation methods and by experimenting with new, previously unavailable training data. The application and tool, Edyson, presented in section 3 below, is an example of the innovative applications that are being developed at KTH in this area.

2.2 Inclusion & Accessibility

Inclusion and Accessibility is a key area for a sustainable and humane future society in a number of policies and strategies. Among these, we find UN Resolution 70/1, "Transforming our World: the 2030 Agenda for

Sustainable Development". The European commission's policy "*Digital Inclusion for a better EU society*" aims to ensure that everybody can contribute to and benefit from the digital economy and society through the development of assistive technologies. The Swedish government has published an action plan for agenda 2030, which again points to inclusion as a key area. Among its directives, it states that people's ability to participate in society shall not be governed by their background, their needs, or their preconditions. Speech technology plays a key role here, in particular with respect to the wide range of people who for whatever reason are excluded from written information.

A current project at KTH develops and makes available Swedish text-to-speech, or speech synthesis, that is suitable for reading aloud the kind of lengthy, complicated texts found in books. Talking books and audio books are notoriously expensive to produce, yet they are essential to inclusion. The project has a special focus on designing and testing innovative evaluation methods for measuring the usability of speech synthesis for the general public and for people with cognitive impairments. The ability to control speaking style, speech rate and articulatory clarity to optimally cater for different users and listening conditions is also a research topic. The project produces speech technology research impact in the *Inclusion and Accessibility* area by paving the way for making all books available as talking books or audio books. It pushes Swedish speech synthesis forward by researching methods specifically targeting the particular difficulties that arise when synthesizing read aloud books, as well as defining currently non-existing evaluation criteria for such speech syntheses.

2.3 Health & Aging

The Europeans Commission's policy "*Research and innovation in digital solutions for health, wellbeing and ageing*" includes both innovating health systems and promoting technology that supports healthy and independent living for the elderly. The range of successful speech technology applications in this area is growing at significant speed. Examples include support systems for the health sector, such as automatic transcription of prescriptions; teaching and training applications; and systems for diagnosis, prevention, treatment and rehabilitation.

The project, "Dialogue for Rehabilition," develops and makes available methods for human-computer dialogues designed to assist with (early) diagnosis, prevention, and rehabilitation. There are a number of health care areas in which relatively repetitive and simple dialogues are used for these purposes (e.g. dementia, autistic spectrum disorders, Parkinson). The project aims to research and generalize these dialogues and to create a dialogue platform that is specifically designed for their implementation. Integrity issues pose a major hurdle here, as they make it difficult, for example, to use commercial cloud based solutions. The project produces speech technology research impact in the *Health & Ageing* impact area by laying the foundation for a kind of patient-machine dialogue that has already proven very efficient in experiments. It pushes Swedish spoken dialogue system research by providing a new clear and useful dialogue type and connecting this to end users, and it will drive research into speech anonymization methods.

2.4 Digitalization

The area of digitalization is highlighted by the European Commission, by the Swedish government, and by KTH's long-term strategy as important for providing new revenue and value-producing opportunities in the process of moving to digital business and customer services. Conversational AI platforms using speech technology are seen to be among the strongest instigators of investments that exploit AI in the near future. This area of speech technology is one of the key strengths of Swedish speech technology in general and of KTH in particular.

At KTH we are building Conversational AI systems that give non-experts easy access to advanced support and guidance. In this project, we explore how conversational systems can be improved through better use of interactional data and by exploring additional modalities such as gaze and breathing. We focus on methods that detect human activities and affective states, and investigate how these can improve conversational skills of e.g. social robots and intelligent voice assistants. The project concerns the digitalization of companies and their processes with the help of speech technology.

3. Edyson

This section describes Edyson, a web-based framework for browsing and annotating large amounts of speech and audio data.

3.1 Temporally disassembled audio

Edyson is based on the notion of temporally disassembled audio (TDA), which is the idea of deconstructing an audio file along its temporal axis with the intent of producing a set of unordered sound snippets of short duration (Fallgren, Malisz, and Edlund, 2019a). Given a set of these short sounds one could rearrange them, and as such listen to them, in any order or manner one wants. Perhaps most importantly, the sounds do not need be ordered in a conventional 1-dimensional sequence, but could for instance be arranged along two axes according to some feature – as is the case of Edyson. The purpose of the process is to remove the time constraints that typically come with analyzing large audio files manually.

3.2 Audio processing

Given an audio file, the audio processing pipeline of Edyson conceptually consists of three main steps. First, the audio is temporally disassembled into snippets of equal length, typically in the range of a few hundred milliseconds to a couple seconds. Second, feature extraction is performed for every snippet, as such representing the short sound as a vector. The decision of what features to use is a tough problem in itself, and depends on the nature of the sound and the purpose of the analysis. MFCCs (Eyben, Wöllmer and Schuller, 2010) are, however, commonly used for speech and are as such used as default in Edyson. Third, the feature vectors are then run through a dimensionality reduction algorithm, e.g. t-SNE (Maaten and Hinton, 2008) self-organizing maps (Kohonen, 1982) that maps every vector to two dimensions - effectively generating a set of xy coordinates for every sound snippet.

3.3 Interface and functionality

The set of coordinates outputted by the audio processing pipeline are visualized in a 2D plot. The distribution of points, and potentially formed clusters, is based on the nature of the feature space, along with whatever properties of the sound the dimensionality reduction algorithm deems most prominent. In other words, two points that are similarly distributed in the plot should also have similar acoustic properties.

The Edyson interface (see Figure 1) has a list of functionalities of which only the essentials will be covered here; for a more extensive list see (Fallgren, Malisz and Edlund, 2019a) or the online documentation[1]. The most important functionality is the listening function, which allows the user to listen to the temporally disassembled audio. This is done by simply hovering over a region of points using the cursor, the system then samples randomly from the selected points and plays the sounds with some overlap which produces a blend of sounds. The parameters of the listening function can be adjusted in real-time in the Edyson interface. For more information on the listening function see Cocktail (Edlund, Gustafson and Beskow, 2010; Fallgren, Malisz and Edlund, 2018). If the user finds an interesting region they can assign a label to it by coloring the points of interest; the timeline then provides instantaneous feedback on where the colored region occurs in the original audio. Furthermore, the dimensionality reduction algorithm can be dynamically changed in real-time which may give the user new information. Edyson can also output any potential findings, specifically, the export function temporally reassembles every sound snippet with their respective label (color). The output can then be imported into other software for further analysis.

3.4 Exploration and annotation

The reason for using Edyson is at least twofold. First, it is an appropriate method for browsing some audio quickly and as such a way for researchers to gain insight into the nature of their data. This is a task that might seem trivial at first, but it is often challenging given the large size of modern audio collections. As an example, the National Library of Sweden in Stockholm hosts many millions of hours of audio-visual data. It is entirely conceivable that a lot, if not most, of these data, are not properly labeled, as is the case for other speech archives and audio collections. The process of TDA allows for fast and efficient browsing of audio which greatly facilitates many downstream tasks within research and audio analysis.

Edyson can also be used for annotation; however, it should be noted that its purpose is not to rival existing software highly specialized for control and efficiency for annotation. Rather, the annotation functionality in Edyson simply serves to provide the user with a basic set of labels of their findings, that for instance could be refined in further analysis.

3.5 Previous results and application areas

Although Edyson is still a work in progress it has shown potential for several different tasks. Fallgren, Malisz, and Edlund (2018) presented an early version and found evidence for the exploration aspect of the method. Specifically, it was shown that the TDA approach could be used to gain insight into different types of audio, e.g. speech, short speech segments, music, and animal sounds. Fallgren, Malisz and Edlund (2019b) conducted experiments where participants explored and properly labeled speech and applause segments in ~10 hours of presidential speeches in a matter of minutes. Most

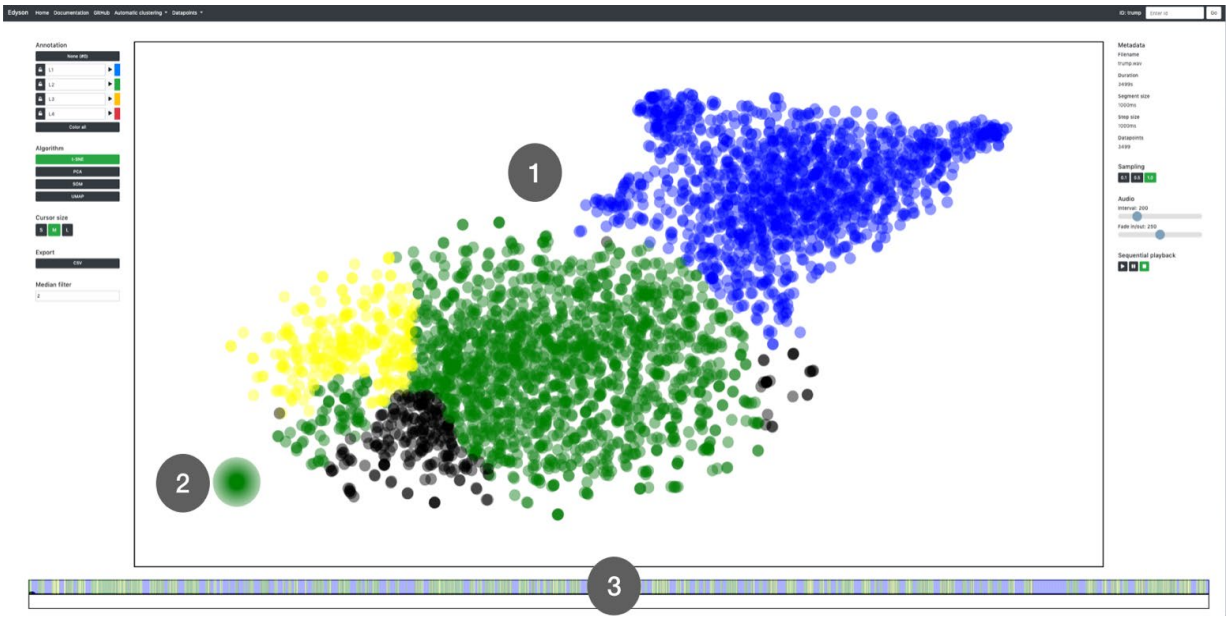


Figure 1 : Edyson interface during browsing of a 1 hour long presidential speech. 1) 2D plot, every point corresponds to a 1 second segment ; 2) Cursor, used for listening and coloring of points ; 3) Timeline, gives instant feedback when coloring.

[1] See github.com/perfall/Edyson for further information and installation instructions.

importantly, the participants did not have any prior knowledge of what the audio contained. Fallgren, Malisz, and Edlund (2019a) provided further evidence for the exploration and annotation aspect of Edyson by annotating 100 hours of noisy radio transmission data for speech activity in less than 45 minutes. Furthermore, pilot results have shown that the tool can be used to gain insights in several areas of speech and audio analysis, e.g. vowel detection, speaker separation, music browsing, noise detection to name a few. Potential application areas are mostly limited by the selection of features, as they carry the information that may or may not capture certain aspects of sound.

Currently Edyson is being used in the project TillTal (Berg et al. 2016), that aims to make cultural heritage recordings accessible for speech research. The Institute for Language and Folklore (ISOF) is engaged in the project and hosts more than 20,000 hours of speech recordings, most of which are digitized. The technology presented here has proven to be a fruitful resource regarding the task of utilizing the large quantities of speech data at hand.

Another reason for using Edyson is that it is completely language independent; one could even explore a recording containing two different spoken languages with the hope of finding similarities or distinctions. It may also help reveal contents of one's data that would otherwise not be found. For instance, when browsing the contents of an hour-long archived interview it was directly evident that there was a minute-long violin-segment in the middle of the recording. In many scenarios observations like this are important and shed light upon the importance of human-in-the-loop frameworks like Edyson.

4. Conclusion

There is currently a wide range of diverse activities in Sweden, particularly at KTH, for research and development of speech technology oriented to the Swedish language. While much of this activity is specifically oriented to Swedish, many of the resulting tools and applications, such as the tool, Edyson, presented here, can be used or adapted for use for any language.

5. Acknowledgements

The work reported on here is supported by the Swedish Research Council (Swe-Clarin and the National Swedish Language Bank, VR 2017-00626) and the Swedish Foundation for the Humanities and Social Sciences (SAF16-0917:1).

6. Bibliographical References

Berg, J., Domeij, R., Edlund, J., Eriksson, G., House, D., Malisz, Z., Nylund Skog, S. and Öqvist, J. (2016). Tilltal – making cultural heritage accessible for speech research. Proceedings CLARIN Annual Conference 26–28 October, Aix-en-Provence, France.

Fallgren, P., Malisz, Z., and Edlund, J. (2018). Bringing order to chaos: a non-sequential approach for browsing large sets of found audio data. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resource Association (ELRA).

Fallgren, P., Malisz, Z., and Edlund, J. (2019a). How to annotate 100 hours in 45 minutes. Proc. Interspeech 2019, pages 341-345.

Fallgren, P., Malisz, Z., and Edlund, J. (2019b). Towards fast browsing of found audio data: 11 presidents. In Proceedings of Digital Humanities in the Nordic Countries, DHN 2019, pages 133-142, Copenhagen, Denmark.

Edlund, J., Gustafson, J., and Beskow, J. (2010). Cocktail–a demonstration of massively multi-component audio environments for illustration and analysis. Proceedings Third Swedish Language Technology Conference (SLTC 2010) page 23, Linköping, Sweden.

Eyben, F., Wöllmer, M., and Schuller, B. (2010). Opensmile: the Munich versatile and fast open-source audio feature extractor. In Proceedings of the 18th ACM International Conference on Multimedia, pages 1459-1462. ACM.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43(1): 59-69.

Maaten, L. V. D., and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research,* 9: 2579-2605.

Proceedings, pages 5–8
Paris, UNESCO Headquarters, 5-6 December, 2019.

Using Language Technologies to Automate the UNDP Rapid Integrated Assessment Mechanism in Serbian

Vuk Batanović *, Boško Nikolić †
* † School of Electrical Engineering, University of Belgrade
* Innovation Center, School of Electrical Engineering, University of Belgrade
Bulevar kralja Aleksandra 73, 11120 Belgrade, Serbia
vuk.batanovic@ic.etf.bg.ac.rs, nbosko@etf.bg.ac.rs

Abstract

Rapid Integrated Assessment (RIA) is a United Nations Development Programme procedure involving a comparison between a country's development policy documents and the UN-defined Sustainable Development Goals. In this paper, we present the Serbian AutoRIA system that automates this procedure in Serbian, a resource-limited yet morphologically rich language. We discuss the issues regarding the preprocessing of data for this task, and the general architecture and language-related specificities of the system. We also evaluate the performance effects of various system settings using the results of a previous, manually completed RIA procedure for Serbia.

Keywords: social good, Sustainable Development Goals, semantic search, word embeddings, word2vec

Résumé

Brza integrisana procena (engl. RIA) je procedura Programa Ujedinjenih nacija za razvoj koja podrazumeva poređenje državnih strateških dokumenata o razvoju i ciljeva održivog razvoja koje su definisale Ujedinjene nacije. U ovom radu predstavljamo srpski AutoRIA sistem koji automatizuje ovu proceduru na srpskom, jeziku sa ograničenim resursima, a razvijenom morfologijom. Razmatramo probleme koji se tiču pretprocesiranja podataka za ovaj zadatak, kao i opštu arhitekturu i jezičke specifičnosti sistema. Takođe evaluiramo efekte različitih podešavanja sistema na njegove performanse koristeći rezultate ranije, ručno sprovedene RIA procedure za Srbiju.

1. Introduction

In September 2015, the United Nations adopted the Agenda for Sustainable Development by 2030, which aims to direct development policies globally. It contains 17 Sustainable Development Goals (SDGs) divided into 169 SDG targets. The Agenda addresses various global challenges, including those related to poverty, inequality, climate, environmental risks, cooperation, and peace and justice. In order to assess a country's readiness for SDG implementation, the UN Development Programme (UNDP) created the Rapid Integrated Assessment procedure (RIA). RIA involves a manual examination of laws, plans, strategies, and other relevant documents, with the aim of determining the degree of alignment between a national development framework and the goals and targets of the 2030 SDG Agenda.

Performing a Rapid Integrated Assessment requires significant human expert labor, and is, thus, costly, in terms of both time and finance. The required human expertise includes not only a high degree of domain knowledge, but also proficiency in the language in which the relevant national documents are written, which may be prohibitive factors for minor languages.

In this paper, we present a system that automates the RIA procedure in Serbian, a morphologically rich yet resource-limited language, using natural language processing (NLP). This work was carried out within the scope of a UN Development Operations Coordination Office innovation project, led by the UN Country Team in Serbia. All the resources and the programming code in Python are publicly available at the *Serbian AutoRIA* GitHub repository[1].

The remainder of the paper is structured as follows: in Section 2, we review previous work on RIA automation, as well as related NLP work for the Serbian language. We also outline a previous, manually completed RIA in Serbian. In Section 3, we discuss the preprocessing of data in Serbian and the architecture of the Serbian AutoRIA system, and in Section 4 we evaluate its results. Section 5 contains our conclusions and some potential avenues of future research.

2. Related Work

The first and, to the best of our knowledge, the only previous effort in automating the RIA procedure was a semantic search approach by Galsurkar et al. (2018). It is centered on a semantic model that compares the meaning of every SDG target description with the meaning of every sentence/paragraph from a policy document. This model was applied to data in English, using previously conducted manual RIAs and national development plans for Bhutan, Cambodia, Liberia, Mauritius, and Namibia.

Previous work on the semantic similarity of short texts in Serbian has been limited. Furlan, Batanović, and Nikolić (2013) proposed a short-text similarity method based on using string and semantic measures with term frequency weighting, and evaluated it on the paraphrase detection task. Batanović, Cvetanović, and Nikolić (2018) presented a corpus of sentence pairs in Serbian annotated with fine-grained similarity scores, and used it to evaluate several supervised and unsupervised semantic similarity models. They found that combining term frequency weighting with a part-of-speech weighting strategy proposed in (Batanović and Bojić, 2015) yielded the best results.

A manual RIA for Serbia was conducted in 2018 by eight policy experts working for close to three months. A total of 145 potentially relevant national documents were detected, but an online document file was found for only 132 of them.

[1] https://github.com/UNDP-Serbia/SerbianAutoRIA

3. Serbian AutoRIA System

Due to the very limited scope and timeframe of the project within which this research was undertaken, we focused mostly on adapting the approach of Galsurkar et al. (2018) to the resource limitations and morphological specificities of Serbian, and verifying its viability in this setting. We will first describe the particularities of preprocessing the data in Serbian, and then the functioning of the AutoRIA system.

3.1 Data Preprocessing

We encountered three main data preprocessing issues – limitations regarding data availability and readability, script variation due to the digraphia present in Serbian, and the high morphological complexity of the language.

3.1.1 Document Collection and Text Extraction

Most of the 132 policy documents collected during the manual RIA process for Serbia were in PDF format, while some were in the form of Word or Excel files. File formats became a major concern in the preparation of data for the AutoRIA system, since NLP models operate on plain texts, and the extractability of plain text from different file formats varies greatly. A specific requirement regarding text extraction was the preservation of correct separation of distinct textual units, such as list items, table cells, headings, paragraphs, etc. This was necessary, since policy documents (e.g. action plans) often contain large tables and lists, and a semantic match for an SDG target can often be found in a single list item or table cell. Similarly, a target match can be a particular document heading or paragraph.

Automatic text extraction from PDFs proved to be quite problematic – no text extraction libraries we experimented with (e.g. *PDFMiner*, *PyPDF2*, etc.) retained the correct text formatting from the original documents. Instead, table structure was lost in the output, with table cells converted out of order, making it impossible to confidently determine the boundaries between cell contents. Paragraphs were broken into multiple text lines, often in the middle of a sentence, depending on how many lines a paragraph was visually separated into within the PDF. The same issue occurred with longer list items, headings, etc. This made it impossible to detect the boundaries between individual textual units in the extracted plain text, preventing the AutoRIA system from functioning properly, as it is based on comparing the meaning of individual textual units with the meaning of each SDG target. Moreover, though most PDF files were generated via Word to PDF conversion, some were actually collated page scans. The low quality of these scans made proper text extraction using optical character recognition infeasible. All these issues prompted us to replace the PDF files with a more suitable file format. We opted for Word files, since they are the most prominent file type after PDFs in the legal domain in Serbia, and since simple copying of their contents into plain text preserves the original text layout as much as possible. We used two main sources for procuring the replacement files:

- www.pravno-informacioni-sistem.rs – the web service of the Official Gazette of the Republic of Serbia
- www.srbija.gov.rs – the official website of the Government of the Republic of Serbia

In cases when these sources were insufficient, we searched the websites of various government ministries, NGOs, etc. However, despite our best efforts, Word file equivalents could not be found for all policy documents. In such cases, we turned to alternative file formats that are still superior to PDF in terms of text extractability, like HTML and Excel. In the end, out of the initial 132 documents, Word files were found for 110, an Excel file for one, HTML files for three, while no replacements were found for the remaining 18 PDFs. Out of those 18, we used *Adobe Acrobat Pro*'s proprietary plain text extraction module on 15, while three documents were discarded due to poor text extractability. We found *Acrobat*'s implementation to be superior to that of open source libraries, though still not perfect, especially with regard to preserving the ordering of table cells and footnotes in the extracted text.

The final policy document set therefore totaled 129 documents converted into plain text. The length of the documents varied widely, from a minimum of around a dozen pages, to a maximum of circa 1500 pages.

3.1.2 Writing Script Normalization

Serbian is a digraphic language with official use of both the Cyrillic and the Latin script, so the policy document set included documents in either script. Moreover, Latin script letters were often found in Cyrillic documents, usually due to a verbatim term from one of the European languages. To avoid a model treating the same word written in different scripts as distinct words, we transliterated all Cyrillic texts into the Latin script using the *CyrTranslit* library[2].

3.1.3 Morphological Normalization

In order to reduce the effects of the morphological complexity of Serbian on data sparsity, a morphological normalizer was required. Previous work on comparing such algorithms for Serbian on various semantic tasks (Batanović and Nikolić, 2017; Batanović, Cvetanović, and Nikolić, 2018) demonstrated that a stemmer for Croatian, a closely related language, by Ljubešić, Boras, and Kubelka (2007), is usually the best-performing option. We therefore applied this stemmer, as implemented in the *SCStemmers* package (Batanović, Nikolić, and Milosavljević, 2016), to all extracted document texts and to the official translations of SDG target definitions and indicators to Serbian.

3.2 System Functioning

Per (Galsurkar et al., 2018), the semantics of an SDG target are represented by the mean of *word2vec* embeddings (Mikolov et al., 2013) of all words in its description. Similarly, the semantic representation of a textual unit (list item, table cell, paragraph, etc.) in a policy document is the mean of the embeddings of words it consists of. Vector scaling based on word TF-IDF scores is also an option. We obtained the embeddings by training the *word2vec* model using the *gensim* library (Řehůřek and Sojka, 2010) on our corpus of 129 policy documents, as Galsurkar et al. (2018) found such specialized embeddings superior to general pre-trained ones when combined with TF-IDF scaling.

Detecting alignments between a policy document and SDG targets is done by measuring the cosine similarity between the semantic vector of each textual unit in the document and the vector of each SDG target. These similarity measurements produce a list of candidate matches for each SDG target, ranked according to their semantic similarity scores. The top-ranking candidate textual units can be given

[2] https://github.com/opendatakosovo/cyrillic-transliteration

over to human experts for closer examination, or compared to existing manual RIA findings for system evaluation.

Galsurkar et al. (2018) proposed two ways of improving this basic system setup by using previously completed manual RIAs. Firstly, they suggested basing the semantics of an SDG target not only on its description, but on a "target document" which also includes all textual matches for that target found in previously completed RIAs. Secondly, they experimented with calculating word TF-IDF scores using the pool of such target documents, instead of the policy document corpus. They found that these changes generally improved system performance on English data, so we decided to evaluate them on Serbian data as well.

4. Evaluation

We first describe the evaluation metrics and data and then present the results of the Serbian AutoRIA system.

4.1 Evaluation Setup

The evaluation metric established by Galsurkar et al. (2018) is the percentage of manually detected SDG target matches that are also chosen as candidate matches by the AutoRIA system. As the number of generated candidates per target increases, this metric will tend to converge to 100%. Since there are numerous SDG targets, the main metric we use is the average percentage of true RIA matches identified by the system, across all SDG targets.

In order to perform this evaluation on Serbian data, we relied on the previously completed, manual RIA for Serbia. The true, manually detected SDG target matches had not been extracted from the documents, so we manually copied the textual units matching the first five SDGs – those under the heading "People". We extracted a total of 342 matches. However, this data was also required to create the aforementioned SDG target documents. Since the manually completed RIA for Serbia was the only such resource in the Serbian language, it was necessary to split the matched textual units into a training set, used for target document creation, and a test set, used for system evaluation.

The data had to be divided in terms of documents, with some of them being placed in the training set, and others in the test set. We tried to maximize the uniformity of SDG target match distribution between the two sets, within the range of a standard training/test split. This, however, presented a rather difficult optimization problem, because a single document often contained matches related to several SDG targets. The matches relevant to SDGs 1–5 occurred in 42 documents. After careful consideration, 31 documents were placed in the training set and 11 in the test set, dividing the data in a 75% – 25% split. Table 1 depicts the resulting distribution of SDG target matches across the two sets. As seen in the table, the overall balance of target matches between the two sets closely follows the document balance. There is some deviation from the overall balance for some SDGs, but it is not excessive and is still within the desirable range for a training/test split.

SDG	Training set	Test set	Total
1	63 (77.78%)	18 (22.22%)	81
2	97 (78.86%)	26 (21.14%)	123
3	30 (65.22%)	16 (34.78%)	46
4	43 (70.49%)	18 (29.51%)	61
5	20 (64.52%)	11 (35.48%)	31
1 – 5	253 (73.98%)	89 (26.02%)	342

Table 1: The distribution of SDG target matches

In our evaluation, the cutoff limit for the number of candidate matches returned for each SDG target was set to 300, per (Galsurkar et al., 2018), to enable some comparability between the results on two different languages. We examined the following system settings:

- Word embedding size – we considered the values of 300, 500, and 1000. Other *word2vec* hyperparameters were set to values used by Galsurkar et al. (2018).
- Using stemming or not.
- Using TF-IDF scaling for word embeddings or not; we also considered calculating TF-IDF scores using the target documents, instead of the policy document corpus.
- Placing SDG target indicators from the 2030 SDG Agenda in their respective target documents or not.
- Placing SDG target matches from the training set in their respective target documents or not.

4.2 Evaluation Results

First, we found that increasing the embedding size does not lead to performance gains, yet augments the computational cost and the time required to complete the analysis. It is thus optimal to use lower-dimensional embeddings, such as the ones with 300 dimensions, for the Serbian AutoRIA.

We then kept the embedding size and the stemming option fixed and considered the effects of other settings, and we plotted and compared the different performances, as shown in Figure 1. System designations are as follows:

- NBOW – the basic model – does not use TF-IDF scaling, SDG target indicators, nor training set target matches.
- TFIDF – a model in which TF-IDF word scaling is used, but SDG target indicators and training set target matches are not. TF-IDF values are calculated using the entire national policy document corpus.
- TFIDF + Ind – the same model as TFIDF, except it uses SDG target indicators.
- TFIDF + TS – the same model as TFIDF, except it uses SDG target matches from the training set.
- Target TFIDF + TS – the same model as TFIDF + TS, but TF-IDF values are based on target documents only.

As seen in Figure 1, TF-IDF scaling generally improves system performance. Placing SDG target indicators into the target documents, however, proves highly detrimental. The probable cause of this effect is that the indicators do not describe the *desired* state of the world (which is the focus of SDG targets), but rather the *current* state. Hence, their use tends to mislead the model into choosing candidate matches that may be related to the semantics of a target, but are not relevant in the narrower context of RIA.

On the other hand, including training set target matches into the target documents is indeed very useful, as claimed by Galsurkar et al. (2018). However, contrary to their findings, calculating TF-IDF scores using target documents, instead of the entire policy document corpus, leads to a performance drop. This divergence between English and Serbian models is likely due to different amounts of data of both types available in each language. The Serbian policy document corpus is over twice the size of the English one, but the target document set is much larger in English, as it includes target matches from five RIAs and for all SDGs. By contrast, the Serbian set is limited to the training portion of target matches from a single RIA for only five SDGs. The quality of TF-IDF weights greatly depends on the size of the corpus used to estimate them, which is why target document-based TF-IDF performs worse on Serbian data

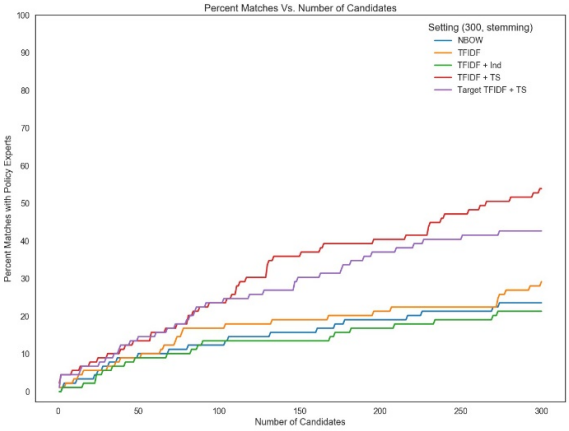


Figure 1: Effects of different settings on system results

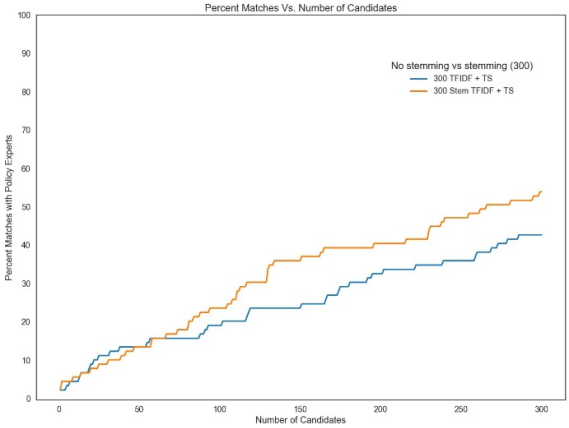


Figure 2: Effects of stemming on system results

than policy document-based TF-IDF.

In order to observe the effects of either using stemming or not, we selected the best-performing models in both settings (in both cases: TFIDF + TS) and we plotted their performances across different cutoff points. Figure 2 shows that on low cutoffs stemming does not necessarily lead to performance improvements, but as the cutoff increases, the benefits of stemming become more pronounced. At the cutoff value of 300, the model that uses stemming outperforms the one that does not by more than 10%.

While the results for the two languages are not directly comparable, this best model for Serbian seems to perform worse at lower cutoff values than the systems designed for English (Galsurkar et al., 2018). Conversely, at higher cutoffs, the system for Serbian is quite similar to the average performance of the models for English, despite being trained with much fewer previous RIA matches.

5. Conclusion

In this paper, we have presented a system for automating the Rapid Integrated Assessment procedure in Serbian, a language with limited resources yet rich morphology. We have discussed the specificities of the Serbian AutoRIA regarding data preprocessing, and the evaluation setup we used to explore the effects of different system settings.

Our findings indicate that, with careful data preprocessing and usage, promising results can be achieved even with scarce manual RIA data. Enlarging the training set would likely lead to even better performance, as would replacing the context-free word embeddings with contextual ones. As is, the current system cannot act as a substitute for human experts, but it can be a tool for assisting them in checking their findings and improving the coverage of their analysis.

6. Acknowledgements

This work was supported by the UNCT project "The use of Artificial Intelligence to automate the Rapid Integrated Assessment mechanism and to nationalize Sustainable Development Goals in Serbia". The authors closely collaborated with the SeConS Development Initiative Group and the staff of UNDP and UNCT Serbia on this project. The authors were also partially supported by the III 44009 research grant of the Ministry of Education, Science and Technological Development of the Republic of Serbia.

7. Bibliographical References

Batanović, V., and Bojić, D. (2015). Using Part-of-Speech Tags as Deep-Syntax Indicators in Determining Short-Text Semantic Similarity. *Computer Science and Information Systems*, 12(1), pp. 1–31.

Batanović, V., Cvetanović, M., and Nikolić, B. (2018). Fine-grained Semantic Textual Similarity for Serbian. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, pp. 1370–1378.

Batanović, V., and Nikolić, B. (2017). Sentiment Classification of Documents in Serbian: The Effects of Morphological Normalization and Word Embeddings. *Telfor Journal*, 9(2), pp. 104–109.

Batanović, V., Nikolić, B., and Milosavljević, M. (2016). Reliable Baselines for Sentiment Analysis in Resource-Limited Languages: The Serbian Movie Review Dataset. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, pp. 2688–2696.

Furlan, B., Batanović, V., and Nikolić, B. (2013). Semantic similarity of short texts in languages with a deficient natural language processing support. *Decision Support Systems*, 55(3), pp. 710–719.

Galsurkar, J., Singh, M., Wu, L., Vempaty, A., Sushkov, M., Iyer, D., Kapto, S., and Varshney, K. R. (2018). Assessing National Development Plans for Alignment with Sustainable Development Goals via Semantic Search. In *30th AAAI Conference on Innovative Applications of Artificial Intelligence (IAAI 2018)*, New Orleans, Louisiana, USA, pp. 7753–7758.

Ljubešić, N., Boras, D., and Kubelka, O. (2007). Retrieving Information in Croatian: Building a Simple and Efficient Rule-Based Stemmer. In *INFuture2007: Digital Information and Heritage*, Zagreb, Croatia, pp. 313–320.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In *Proceedings of the International Conference on Learning Representations Workshop (ICLR 2013)*, Scottsdale, Arizona, USA.

Řehůřek, R., and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, Valletta, Malta, pp. 45–50.

Proceedings, pages 9–12
Paris, UNESCO Headquarters, 5-6 December, 2019.

CALL Solutions that Support Multilingualism: Application to the "Nano" Languages in the West-Nordic Region

Birna Arnbjörnsdóttir, Auður Hauksdóttir
Vigdís Finnbogadóttir Institute, University of Iceland
birnaarn@hi.is, auhau@hi.is

Abstract

This article describes innovative CALL solutions that support multilingualism and "nano" languages in the West-Nordic Region. The non-language specific platforms are developed at the Vigdís Finnbogadóttir Institute at the University of Iceland and include open curated language courses for newcomers, www.icelandiconline.com, www.faroeseonline.fo and tools that enhance oral fluency and communicative skills in Danish as a second and third language in the West-Nordic region. The tools include: www.talboblen.hi.is, that focuses on oral language skills, www.talerum.is , an interactive game based program that promotes interaction, and www.frasar.net, a resource that teaches the pragmatics of phrases.

Keywords: CALL, Language Education, Linguistic Diversity, Second Language Acquisition.

Útdráttur

Í greininni er fjallað um nýstárleg tölvustudd námsgögn, sem er ætlað að stuðla að fjöltyngi og vexti og viðgangi tungumála í örsmáum málsamfélögum á Vestur-Norðurlöndum, þ.e. í Færeyjum, Grænlandi og á Íslandi. Námsgögnin hafa verið þróuð af starfsmönnum Stofnunar Vigdísar Finnbogadóttur við Háskóla Íslands, en sú þekking, sem hefur skapast, getur gagnast hvaða tungumáli sem er. Annars vegar er um að ræða opin vefnámskeið í íslensku - www.icelandiconline.com - og færeysku - www.faroese.fo - og hins vegar þrjú máltæki, sem eru hönnuð með tilliti til kennslu dönsku sem erlends máls, og er ætlað að stuðla að tjáskiptahæfni nemenda, einkum hvað varðar talmál. Máltækin eru www.taleoblen.hi.is, þar sem megináhersla er lögð á talmálsþjálfun, gagnvirki tölvuleikurinn www.talerum.is, sem reynir á notkun dönsku í ólíkum málaðstæðum, og www.frasar.is, sem kennir orðasambönd í dönsku og íslensku.

1. Introduction

The linguistic communities of the West-Nordic Region are characterized by their diversity. The region includes four countries: The Faroe Islands, Iceland, and Greenland and Norway. All are part of The Nordic Countries or "Norden", which include Denmark, Sweden, Finland, Norway and Iceland as well as the autonomous territories of the Faroe Islands, Greenland and Åland, plus the Saami regions. A total of over 27 million people live in the Nordic countries today (Hauksdóttir, Lund & Börestam 2016: 93). The languages spoken in this area belong to three different language families:

1. Germanic languages include Danish, Faroese, Icelandic, Norwegian (*bokmål* and *nynorsk*) and Swedish,
2. Finnish and Saami are Finno-Ugric languages,
3. Greenlandic is classified as an Eskimo-Aleut language.

The Nordic region also comprises multilingual communities, where some 200 non-Nordic languages are spoken (Declaration on a Nordic Language Policy 2007:11). In this article we focus on efforts to support the local languages spoken in Iceland, the Faroe Islands and Greenland, and those who learn them as second languages and efforts use to support multilingualism in the these changing linguistic communities whose speakers must increasingly seek information, education and communication with the outside world in languages that are not their first language; in most cases Danish and/or English. These speech communities have in common that the local languages have very few speakers. Greenlandic is spoken by approximately 55.000 speakers, Faroese has about 75.000 speakers and Icelandic is spoken by around 320.000 speakers (Norden i skolan).

The three languages which can be described as "nano" languages are the majority and official languages of the respective country and used in education and governance. Official efforts are being mounted to sustain the languages as first languages that cover all domains of language and at the same time support newcomers in these speech communities to learn the local language quickly and with no cost. Supporting sustainability is important to prevent language attrition and possibly language loss (Fishman 1996, Haugen 1972).

This article describes five Computer Assisted Language Learning (CALL) projects developed at the Vigdís Finnbogadóttir Institute of Foreign Languages at the University of Iceland. Two of the projects are aimed at newcomers learning the local languages of Iceland and the Faroe Islands, while three projects support Danish as a second or third language instruction in the West Nordic region. All the systems can be used to teach other

languages.

2. Background

Greenland has two official languages, Greenlandic and Danish while in the Faroe Islands and Iceland, Faroese and Icelandic are the sole official languages. Faroese, Icelandic and Greenlandic have for centuries been in intense contact with Danish. Politically, the Faroe Islands and Greenland are part of the Danish Monarchy while Iceland became an independent republic in 1944. Danish influence has thus been robust in all three communities although this is seen as decreasing. Today, English occupies an important and growing role in the linguistic landscape of all three speech communities. In all three countries, English has an official status as a foreign language. Danish is considered a foreign language in the Faroe Islands and Iceland although the presence of Danish is more evident in the former. Greenlandic and Faroese have several recognized geographical varieties while in Iceland variation is minimal.

The three countries are highly dependent on foreign relations. Proficiency in foreign languages is therefore vital. Language skills are the key to all communication with the outside world and a prerequisite for study abroad and more recently for study at local universities where a majority of textbooks are written in foreign languages, either in Danish (Hauksdóttir 2012b) or in English (Arnbjörnsdóttir & Ingvarsdóttir, 2018).

Danish is a mandatory foreign language subject at the primary school level in the Faroe Islands, Greenland and Iceland, however the actual role of Danish in these countries varies greatly. Due to their links with Denmark, Danish plays a key role in Faroese and Greenlandic society, while it is seldom used in everyday life in Iceland. Danish is the first foreign language taught in Greenland and on the Faroe Islands, but as a second foreign language after English in Iceland. While Danish is taught as a foreign language in secondary schools in the three countries, the volume and goal of instruction varies. The stated main purpose of Danish instruction in curriculum guidelines is to enable learners to participate in the Nordic linguistic and cultural community that encompasses more than 27 million speakers of related languages who share a common cultural heritage (Declaration on a Nordic Language Policy 2007:12). Teaching English is a priority in the Nordic school system, both at the primary and secondary school level. Increasingly Nordic universities are offering programs for local students where English is a Medium of Instruction and most textbooks at university level are in English. The same is true for continuing education and professional development. Without English it is difficult to obtain advanced education in the three countries nor in the Nordic countries in general (Arnbjörnsdóttir & Ingvarsdóttir, 2018) The need for proficient users of foreign languages is seen as significant for individuals and for the possibilities of these countries to make their voices heard in an international context.

The global spread of English through media and popular culture has found its way to the West-Nordic region. Additionally, mobility related to educational and work opportunities has increased the use of English as a Lingua Franca in many areas where previously Danish served the same purpose. The changing yet important role of Danish and English in these countries has put pressure on the local languages. For language communities with few speakers it is important to enhance instructional material in the local language for newcomers, and to support students who must pursue their education in a language other than their first and the one they used in primary school.

3. The CALL Projects

One of the main challenges for linguistic communities with few speakers is the cost of developing adequate instructional materials that meet modern standards. Publishing costs are high and unsustainable in a very small market such as in the West Nordic region. In many ways teaching languages through computers or mobile devices alleviates the need for printed materials. Computer assisted language learning can also serve to motivate a new generation of digital natives or learners who have grown up with computers and the internet.

Web based language teaching offers excellent solutions for small and less commonly taught languages. The *Icelandic Online* platform and pedagogy are based on the notion that adults who wish to learn languages, should be able to do so at their own pace and on their own time. They also benefit from material that has been curated for them especially, to save them the time and effort to find the materials for themselves on the internet. The development of the three tools for supporting Danish described in this article was guided by the same principles. Motivation serves as key as the tools are game-based and use speech recognition to enable individuals to work on their own language development at their own pace. Because the tools cater to the individual, they can be used across linguistic and geographical boundaries where Danish serves different roles and extramural exposure varies (Hauksdóttir, 2004).

3.1 Teaching Icelandic and Faroese Online

In order to support newcomers residing, working and studying in Iceland and the Faroe Islands, the Vigdís Finnbogadóttir Institute has supported the development of www.icelandiconline.com and www.faroese.fo. Both are based on the same technology and pedagogy for teaching the two different languages.

Icelandic Online.com has had over 270.000 visitors from around the world with over 70.000 active users. About 25% of users have IP numbers in Iceland. The six free and open courses are structured, curated and based on a specific pedagogy. A non-language specific course editor can choose from over forty preprogrammed templates that allow course developers to construct courses for different purposes and target groups without the involvement of a programmer. The courses have an in-built tracking system that monitors students' progression through the system. The courses are offered on a multiplatform system that includes smartphones and tablets.

Course content is based on a specific pedagogy that allows the learner to develop accuracy and fluency gradually. This

is a skill-based course with emphasis on lexical, grammatical and pragmatic skills.
www.Icelandic Online.com includes 6 courses. IOL 1 was launched in 2004, IOL 2 l in 2005, IOL PLUS Tutor in 2006 blended, IOL 3 and 4 and IOL survival course for newcomers in 2010 and IOL 5 in 2012.

The pedagogy used in *Icelandic Online* is based on relevant second language theories (Chapelle, 2001). The basic premise is that adults learn language differently than children and that they benefit from instruction, need accuracy practice, have experience learning foreign languages, have divers learning styles, and that are motivated by interesting and relevant content to stay on the course.

Computer Assisted Languages Learning (CALL) is often said to be good for teaching accuracy, but not necessarily fluency. This means that using computers can help learners learn and understand target like language but is less efficient in teaching speaking and writing. CALL is therefore important for teaching morphologically complex languages (Icelandic, Faroese and Greenlandic) especially to adults. Complex morphology poses certain problems for adults who are beyond the optimal age for language development. In the case of Icelandic, nouns and adjectives have many different forms and agree in gender, number and case that are marked by different surface forms. The challenge for CALL is to present the necessary morphology up front in online beginner courses and still retain learners (Arnbjörnsdóttir, 2004).

Before programming a course, the development process begins with the exact goal of the specific target group and language and the ways in which these goals can be met. The interactive pedagogy is adapted to each language proficiency level. Survival courses focus on pragmatics based on the reality in which immigrants in Iceland and the Faroes find themselves; advanced beginner courses have very highly structured and scaffolded materials with glossaries; IOL 3 and 4 have longer texts - written and oral based on every-day life in Iceland. Scaffolding is gradually removed and more emphasis is placed on authentic texts including expository texts, IOL 5 focuses on Icelandic literature.

The European Benchmarks determine proficiency levels: IOL Survival (A1), IOL 1 and 2 A1 - A2) and IOL 3 and 4 (B1 - B2), while IOL 5 is at the C1 level (CEFR).

Icelandic Online's technology and pedagogy are thoroughly tested and has received very positive feedback from users and teachers. IOL is currently offered in three different modes of delivery: Open online courses, blended learning and distance learning with an online tutor. IOL is used by lecturers of Icelandic abroad, in secondary schools, by immigrants, and heritage speakers. Approximately 1/3 of those who enrolled are active learners and gender division is equal. Most of the learners are University students - and most of them come from 10 countries. But learners also come as far away as Burkina Faso and Malaysia. Of the current users, 69% are under the age of 31 years.

IOL has been tracking learners since 2006. The tracking system logs the position of the user as he or she navigates through a course and then creates text files that are stored on a server. Findings indicate that when measured from the 15% juncture in the course rather than from the beginning, retention is high, and few students drop out. Results show clearly that the mode of delivery matters in retention – where the blended mode is most effective. The tracking system revealed signs of dropout bundles at the same specific junctures in all the courses. Upon comparing these junctures to the course content, preliminary indications are that a "storyline" in a course may be important in retaining learners and that assignment submission may both explain why students stay and leave (Friðriksdóttir & Arnbjörnsdóttir, 2015).

3.2 Teaching Danish through CALL in the West-Nordic Region

The lexicon of a language includes lexical phrases such as idioms, collocations and pragmatic phrases (Nattinger & DeCarrico, 1992; Wray, 2002). The website *www.frasar.net* is developed with Icelandic learners of Danish in mind. It contains 7560 idioms and pragmatic Danish phrases and their equivalent in Icelandic, when an equivalent exists. The tool offers different search options, including a whole or part of a phrase, its meaning as well as pragmatic information. Information on the meaning of the phrases can be obtained in both Icelandic and Danish, explaining variation in form and inflection. Idioms are sometimes problematic for foreign language learners, since their meaning typically is not transparent, or their usage can be specific and even at odds with more common use of the words in the target language. Due to their opacity, idioms can be difficult to understand, and due to their unpredictability, language learners are often unsure of their form. Sometimes the same idioms are used in many different languages for the same or similar meaning. This is especially true for related languages, such as Danish and Icelandic, but they often differ in unpredictable ways. Pragmatic phrases can also be relevant and necessary to certain situations of spoken and written language and are of great importance for target like use of language. For example, while greeting formally or informally, making requests or asking for assistance in a shop or in other everyday circumstances. Because of the similarity of the two languages, appropriate use of idioms and pragmatic phrases can be confusing for Icelandic students. In order to make *www.frasar.net* as practical as possible, the use of the phrases can be examined in context both with constructed examples and as they appear in a large corpus on which the tool is based (Hauksdóttir 2012a, 2014a).

Taleboblen is partly based on www.frasar.net, and contains over one thousand common phrases from a corpus of Danish everyday speech with the equivalent phrase in Greenlandic, Icelandic and Faroese. If there is no equivalent, an example and explanation of the Danish phrase is provided with an audio example. *Taleboblen* offers a variety of exercises for practicing oral proficiency and pragmatics in certain linguistic situations such as on the train or bus. Students listen to the pronunciation of Danish lexical phrases and they can record their own pronunciation of the words and phrases and then listen to their own voice. *Taleboblen* includes a tool for practicing intonation and rhythm in pronouncing phrases. Students receive immediate feedback on their pronunciation and their progress (Henrichsen & Hauksdóttir, 2015).

Talerum is a conversation-based game for learning Danish. It is aimed at students in the upper level of primary education in the West Nordic region. The game can also be used to teach Danish as a second language. The learner/user is an exchange student within the game and has a Danish host family that picks him up at the airport at the start of the game. The game includes a variety of tasks or challenges including going to the store to buy food or clothing. The students advance in the game only by using the language to perform the tasks they are assigned. There are many different scenarios in the game and the user's task is to solve puzzles gleaning information from conversations and thereby navigating through to the end of the game. Each scene has a theme e.g. Dagligstuen (e. Living room). Some scenes are independent, and it is not necessary to finish all the scenes in a specific order. But there are scenes that need to be finished before others. In the conversation scenarios, the user is prompted with a question. To answer the question correctly the user has to include one of many accepted keywords for that particular question in his answer. If a question is not correctly answered, the user is asked another question. When a user finally includes a correct keyword, the scene changes and the learner advances to the next set of questions. A speech recognition tool enables learners to hear what is being written in the game. The speech recognition tool is still in development, and *Talerum* will be Beta tested in schools in the West-Nordic region in early 2020.

4. Conclusion

The platforms upon which the projects described in this article are based are specifically developed for teaching language but are not language specific. They are not content management systems. They are also built on a specially developed pedagogy appropriate for online language learning. The technology and pedagogy can serve any language and any course. The positive reviews that the courses have received attest to this. Future projects include Icelandic Online for children in the beginning stages of second language literacy.

5. Acknowledgements

The authors wish to thank their major donors, NORDPLUS, The University of Iceland and The Vigdís Finnbogadóttir Institute, for their financial support for the projects presented in this article.

6. Bibliographical References

Arnbjörnsdóttir, B. (2004). Teaching Morphologically Complex Languages Online: Theoretical Questions and Practical Answers. In Peter Juul Hendrichsen (Ed.) CALL for the Nordic languages. Tools and methods for Computer Assisted Language Learning. Copenhagen Studies in Language, 30, pages 59-74. Copenhagen: Samfundslitteratur.

Arnbjörnsdóttir, B. & Ingvarsdóttir, H. (2018). Language development across the lifespan: The impact of English on education and work in Iceland. Amsterdam: Springer.

CEFR. https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions

Chapelle, C. (2001). Computer applications in second language acquisition: Foundations for teaching, testing, and research. Cambridge: Cambridge University Press.

Declaration on a Nordic Language Policy (2007) : http://norden.divaportal.org/smash/record.jsf?pid=diva2%3A700895&dswid=5408

Friðriksdóttir, K., & Arnbjörnsdóttir, B. (2015). Tracking student retention in open online courses. In F. Helm, L. Bradley, M. Guarda, & S. Thouësny (Eds.), *Critical CALL* – Proceedings of the 2015 EUROCALL conference, Padova, Italy, pages 192-197. Dublin: Research-publishing.net: http://dx.doi.org/10.14705/rpnet.2015.000332

Fishman, J. (1996). What do you lose when you lose your language. In Cantoni G. (Ed.) Stablizing indigenous languages. pages 186–196. Flagstaff: Northern Arizona University Press,

Haugen, E. (1972*). The ecology of language: Essays by Einar Haugen. S*elected and Introduced by Anwar S. Dil. Standford: Standford University Press.

Hauksdóttir, A. (2004). CALL for Communicative Competence in foreign languages. In Peter Juel Henrichsen (Ed.). CALL for the Nordic languages. Tools and methods for Computer Assisted Language Learning. Copenhagen Studies in Language, 30, pages 9–31. Copenhagen: Samfundslitteratur.

Hauksdóttir, A. (2012a). At komme til orde på et mundret dansk. Om fraser, fraseindlærning og fraseværktøj anskuet kontrastivt. Speech in Action. Proceedings of the 1st SJUSK Conference on Contemporary Speech Habits. Copenhagen Studies in Language 42, pages 123–142.

Hauksdóttir, A. (2012b). *Dansk som fremmedsprog i en akademisk kontekst: Om islændinges behov for danskkundskaber under videreuddannelse i Danmark*. Copenhagen Studies in Bilingualism, no. 68. Copenhagen: University of Copenhagen.

Hauksdóttir, A. (2014). Sprogværktøjet: www.frasar.net: Om fraser og fraseindlæring anskuet kontrastivt. *Språk i Norden*, pages 68–82.

Henrichsen, P. J. og Hauksdóttir, A. (2015). Taleboblen – den tålmodige transnordiske udtaletræner. Dorthe Duncker, Eva Skafte Jensen og Ole Ravnholt (Eds.): *Rette ord. Festskrift til Sabine Kirchmeier-Andersen i anledning af 60-årsdagen*, pages 59–170. Copenhagen: Dansk Sprognævns skrifter 46.

Hauksdóttir, A., Lund, J., Börestam, U. (2016). Language and culture link us together. In Debra L. Cagan (Ed.) *Nordic ways*, pages 91–99. Washington: The Johns Hopkins University, Center for Transatlantic Relations.

Nattinger, J. R., DeCarrico, J. S. (1992*). Lexical phrases and language teaching*. Oxford: Oxford University Press.

Norden I skolan (2019). https://nordeniskolen.org/da/sprog-kultur/gymnasiet/de-nordiske-sprog/om-faeroesk/.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Proceedings, pages 13–15
Paris, UNESCO Headquarters, 5-6 December, 2019.

Linguistic Linked Open Data for All

John McCrae[1], Thierry Declerck[2]
[1]Insight Centre for Data Analytics at the National University of Ireland Galway, Ireland
[2]DFKI GmbH, Multilinguality and Language Technology, Germany
[1] john.mccrae@insight-centre.org
[2]declerck@dfki.de

Abstract

In this paper we briefly describe the European H2020 project "Prêt-à-LLOD" ('Ready-to-use Multilingual Linked Language Data for Knowledge Services across Sectors'). This project aims to increase the uptake of language technologies by exploiting the combination of linked data and language technologies, that is Linguistic Linked Open Data (LLOD), to create ready-to-use multilingual data. Prêt-à-LLOD aims to achieve this by creating a new methodology for building data value chains applicable to a wide-range of sectors and applications and based around language resources and language technologies that can be integrated by means of semantic technologies, in particular the usage of the LLOD.

Keywords: Linguistic Linked Open Data, Standards, Infrastructure

Résumé

Dans cet article, nous décrivons brièvement le projet Européen «Prêt-à-LLOD» («Données multilingues prêt à l'emploi pour les services de la connaissance dans tous les secteurs»). Ce projet vise à accroître l'utilisation des technologies langagières en exploitant la combinaison de données liées et de technologies langagières, à savoir les données linguistiques ouvertes et liées (LLOD), pour créer des données multilingues prêtes à l'emploi. Prêt-à-LLOD vise à atteindre cet objectif en créant une nouvelle méthodologie pour construire des chaînes de valeur de données applicables à un large éventail de secteurs et d'applications et reposant sur des ressources linguistiques et des technologies langagières pouvant être intégrées au moyen de technologies sémantiques, en particulier l'utilisation du LLOD

1. Introduction

Language technologies increasingly rely on large amounts of data and better access and usage of language resources will enable to provide multilingual solutions that support the further development of language technologies in Europe and in the world. However, language data is rarely 'ready-to-use' and language technology specialists spend over 80% of their time on cleaning, organizing and collecting language datasets. Reducing this effort promises huge cost savings for all sectors where language technologies are required. An essential part of the Extract-Transform-Load process currently needed involves linking datasets to existing schemas, yet few specialists take advantage of linked data technologies to perform this task. The Prêt-à-LLOD project[1] aims at increasing the uptake of language technologies by exploiting the combination of linked data and language technologies, that is Linguistic Linked Open Data (LLOD), to create ready-to-use multilingual data. Prêt-à-LLOD aims to achieve this by creating a new methodology for building data value chains applicable to a wide-range of sectors and applications and based around language resources and language technologies that can be integrated by means of semantic technologies. The project develops novel tools for the discovery, transformation and linking of datasets and apply these to both data and metadata in order to provide multi-portal access to heterogeneous data repositories. A goal is also to automatically analyse licenses in order to deduce how data may be lawfully used and sold by language resource providers. Finally, the project provides tools to combine language services and resources into complex pipelines by use of semantic technologies. This leads to sustainable data offers and services that can be deployed to many platforms, including as-yet-unknown platforms, and can be self-described with linked data semantics. Our approach is being validated in four pilots.

In the following sections we present briefly the Linguistic Linked Open Data cloud and the OntoLex-Lemon representation model for lexical data, two of the main initiatives upon which Prêt-à-LLOD is building on. We then discuss briefly some of the objectives of the project methodologies put in place in order to reach them, showing also their relevance for less-resourced languages.

2. Linguistic Linked Open Data Cloud

The Linguistic Linked Open Data (LLOD) cloud[2] is an initiative, which was started in 2012 by a group of the Open Knowledge Foundation[3]. The aim was to break the data silos of linguistic data and thus encourage NLP applications that can use data from multiple languages, modalities (e.g., lexicon, corpora, etc.) and develop novel algorithms. Looking at the current state of the LLOD, displayed in Figure 1, one can see that the data sets published in this cloud are classified along the lines of six categories:

- Corpora
- Terminologies, Thesauri and Knowledge Bases
- Lexicons and Dictionaries

[1] https://www.pret-a-llod.eu/

[2] See https://linguistic-lod.org/llod-cloud for more detail.

[3] See (McCrae et al., 2016) for a description of the development of the LLOD.

- Linguistic Resource Metadata
- Linguistic Data Categories
- Typological Databases

Not all the date sets are equally linked to each other, and our project can contribute in better linking the data sets in the fields of Terminologies, Thesauri and Knowledge Bases and those in the fields of Lexicons and Dictionaries.

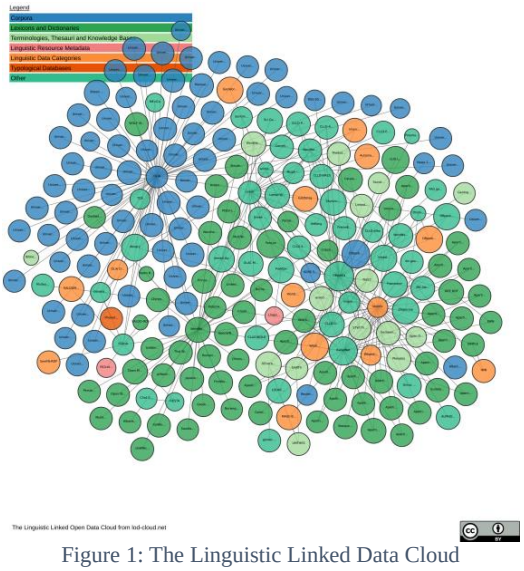


Figure 1: The Linguistic Linked Data Cloud

3. OntoLex-Lemon

The OntoLex-Lemon model, which is resulting from a W3C Community Group[4], was originally developed with the aim to provide a rich linguistic grounding for ontologies, meaning that the natural language expressions used in the labels, definitions or comments of ontology elements are equipped with an extensive linguistic description.[5] This rich linguistic grounding includes the representation of morphological and syntactic properties of lexical entries as well as the syntax-semantics interface, i.e. the meaning of these lexical entres with respect to an ontology or to specialized vocabularies.

The main organizing unit for those linguistic descriptions is the lexical entry, which enables the representation of morphological patterns for each entry (a multi word expression, a word or an affix). The connection of a lexical entry to an ontological entity is marked mainly by the denotes property or is mediated by the *LexicalSense* or the *LexicalConcept* classes, as this is represented in Figure 2, which displays the core module of the model. OntoLex-Lemon builds on and extends the lemon model (Cimiano et al. (2016)). A major difference is that OntoLex-Lemon includes an explicit way to encode conceptual hierarchies, using the SKOS[6] standard. As can be seen in Figure 2, lexical entries can be linked, via the *ontolex:evokes* property, to such SKOS concepts, which can represent WordNet synsets. This structure is paralleling the relation between lexical entries and ontological resources, which is implemented either directly by the *ontolex:reference* property or mediated by the instances of the *ontolex:LexicalSense* class.

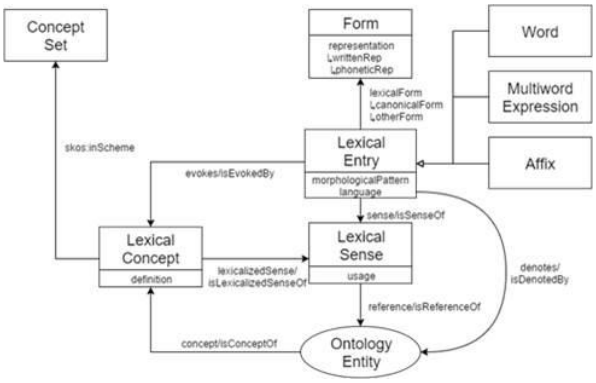


Figure 2: The core Modules of OntoLex-Lemon.
Graphic taken from https://www.w3.org/2016/05/ontolex/.

More recently, OntoLex-Lemon has been used also as a de-facto standard in the field of digital lexicography and is being applied for example in the European infrastructure project ELEXIS (European Lexicographic Infrastructure)[7].

4. Main Objectives of Prêt-à-LLOD

The first goal of this project is to allow for multilingual cross-sectoral data access that supports the rapid development of applications and services to be deployed in multilingual cross-border situations. This is realised by providing data discovery tools based on metadata aggregated from multiple sources, methodologies for describing the licenses of data and services, and tools to deduce the possible licenses of a resource produced after a complex pipeline.

A second goal consists in developing a new ecosystem to support the development of novel linked data-aware language technologies, from basic tools such as taggers to full applications such as machine translation systems or chatbots, based on semantic technologies that have been developed for LLOD to provide interoperable pipelines. We apply state-of-the-art semantic linking technologies in order to provide semi-automatic integration of language services in the cloud

A third goal is concerning sustainability. The sustainability of language technologies and resources is a major concern. We aim to solve this by providing services as data , that is, wrapping services in portable containers that can be shared as single files. Language data also eventually becomes valueless as the documentation and expertise for processing esoteric formats is lost, and apply the paradigm of data as services, where services can be embedded in multi-service workflows, that demonstrates the service's value and

[4] See https://www.w3.org/2016/05/ontolex/

[5] See (McCrae et al., 2012), (Cimiano et al., 2016)

[6] SKOS stands for "Simple Knowledge Organization System". SKOS provides "a model for expressing the basic structure and content of concept schemes such as thesauri, classification schemes, subject heading lists, taxonomies, folksonomies, and other similar types of controlled vocabulary" (https://www.w3.org/TR/skos-primer/).

[7] See http://www.elex.is/ for more detail.

supports long-term maintenance through methods such as open source software. Furthermore, we build supporting tools to measure and analyse the validity, maintainability and licensing of the data and services. This increases the quality and coverage of language resources and technologies by ensuring that services are easier to archive and reuse, and thus remain available for longer. In particular, this goal is important for minority and under-resourced languages, where the precious effort to develop resources is often lost.

5. Methods

The project implements methods for discovery, transforming and linking linguistic data so that they can be published in the LLOD.

Prêt-à-LLOD provide a flexible discovery method that can search over both language resources and services. As many real challenges can only be handled by a combination of multiple datasets and services, the project develops a new workflow system that supports chaining of multiple services using semantic service descriptions and containerization to avoid becoming a "walled garden" ecosystem.

A key challenge for this is the chaining of services and data from heterogeneous sources. To this end, we apply linking to develop a transformation component which uses a novel three-step process whereby data from multiple sources is combined by means of RDF (Resource Description Framework, the representation language needed to publish data in the LLOD, linked and then harmonized using semantic and language technologies. The resulting discovery and search platform consists in a single and user-friendly portal.

An integrated methodology has been designed for the transformation of language resources. The goal of the transformation is either OntoLex-Lemon model (briefly introduced above) for lexical data or any RDF vocabulary supporting the representation of language data. This is an important aspect for less- or under-resourced languages, as they have the same "representation dignity" as other languages, to which they can be linked to, in the LLOD ecosystem.

Finally, the project is developing (semi-)automated linking mechanisms. This concerns both conceptual level of language descriptions as also the lexical data. We are working both in a mono- and in a cross-lingual set up.

As stated above, Prêt-à-LLOD is also concerned with the issue of detecting and "chaining" licensing conditions for the language resources and services that can be combined in complex pipelines. So that additionally to the three basic methodologies described just above, the project is dealing with the automated execution of smart policies for language data transactions.

All those steps need for sure to be carefully designed and integrated in a workflow. Prêt-à-LLOD is therefore designing a protocol, based on semantic markup, that is aiming at enabling language services to be easily connected into multi-server workflows.

6. Standards

Prêt-à-LLOD members are involved in a series of standardisation activities, mainly related to the de-facto standard OntoLex-Lemon. We would like to mention here the new module on lexicography ("lexicog")[8], on a more precise description of morphological phenomena[9] and on the topics of FRequency, Attestations and Corpus data ("frac")[10]. Those new modules are extending the expressive power of OntoLex-Lemon, and also very important for the inclusion of less- and under-resourced language data in the LLOD, as the scope of the de-facto standard is extended to corpus data, besides the coverage of lexical data.

7. Conclusion

We presented the current state of the Prêt-à-LLOD project, which is aiming at further extending the Linguistic Linked Open Data cloud infrastructure and making more language data interoperable, also with sustainable semantic description approaches.

8. Acknowledgements

The project Prêt-à-LLOD has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 825182.

9. References

Cimiano, Philipp, McCrae, John and Paul Buitelaar. 2016. Lexicon Model for Ontologies: W3C Community Report.

McCrae, John, Aguado-de Cea, Guadalupe, Buitelaar, Paul, Cimiano, Philipp, Declerck, Thierry, Gomez-Perez, Asuncion, Garcia, Jorge, Hollink, Laura, Montiel-Ponsoda, Elena, Spohr, Dennis and Wunner, Tobias. 2012. Interchanging Lexical Resources on the Semantic Web. Language Resources and Evaluation, 46(4):701–719

John P. McCrae, Christian Chiarcos, Francis Bond, Philipp Cimiano, Thierry Declerck, Gerard de Melo, Jorge Gracia, Sebastian Hellmann, Bettina Klimek, Steven Moran, Petya Osenova, Antonio Pareja-Lora, Jonathan Pool. The Open LinguisticsWorking Group: Developing the Linguistic Linked Open Data Cloud Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), Portoroˌ, Slovenia, ELRA, ELRA, 9, rue des Cordelières, 75013 Paris, 5/2016

Rodiguez-Doncel,Victor, Casanovas, Pompeu. 2018. A Linked Data Terminology for Copy-right Based on Ontolex-Lemon: AICOL. International Workshops 2015-2017. 410-423.

[8] See https://www.w3.org/2019/09/lexicog/

[9] https://www.w3.org/community/ontolex/wiki/Morphology

[10] https://github.com/acoli-repo/ontolex-frac

Proceedings, pages 16–19
Paris, UNESCO Headquarters, 5-6 December, 2019.

Multilingual Natural Language Processing and Transformers: A Giant Step Forward

Radu Florian Taesun Moon Parul Awasthy Jian Ni
IBM Research AI
Yorktown Heights, NY 10598
{raduf,tsmoon,awasthyp,nij}@us.ibm.com

Recent developments in deep learning for natural language processing have opened up opportunities to develop tools and libraries for multiple languages simultaneously and also for low resource languages. Here we describe these advances as well as our experiments that show that one can build a multilingual named entity recognition system that works well on multiple languages, in addition to being able handle unseen languages.

1. Introduction

Natural language processing is a scientific discipline as well as a field of software engineering which provides a structured, statistical interface to written human language. A well-known, well-established and early application of this field is machine translation where the goal is to translate text in one language to another via a computer program without any human intervention. A more recent application with many commercial and scientific uses is sentiment analysis, which detects whether some statement or review is favorable, unfavorable or neutral toward some subject with many different shades of granularity in terms of both subject matter and sentiment.

An application that lies more in the background but is no less important is information extraction which comprises named entity recognition, relation extraction and coreference resolution. Named entity recognition detects the proper nouns in texts such as "The *UN* will hold its *General Assembly* in *New York* soon. *It* is expected to increase traffic significantly in *Midtown*" where the named entities (and one pronominal mention) are in italics. Relation extraction labels any relations that may exist between such named entities such as the fact that the *General Assembly* will be in a valid textual relation with *New York* that we will label here *to take place in*. Coreference resolution is the component which links together all the textual mentions in a text that refer to the same entity, in this case *General Assembly* and *It*. These three components are usually packaged together and is called an information extraction system, with wide uses in many areas that require a structured or simplified representation of large, unruly natural language corpora such that it can be processed uniformly and quickly by downstream applications such as databases, text analytics engines and automated decision making.

All the above applications require a substantial amount of text that humans have labeled with appropriate information so that the underlying statistical models used by the NLP components have idealized output that it may adopt without having a human engineer manually encode millions or possibly billions of individual behaviors. The time and cost involved in creating this labeled data is still considerable and beyond the reach of most language communities outside a handful of the most commonly used languages such as English and Chinese. Two recent developments in deep learning for NLP give us reason to hope that the pipeline for creating tools for low resource languages is about to be greatly both simplified and improved at the same time. Namely, these are the transformer architecture (Vaswani et al., 2017) and multilingual Bidirectional Encoder Representations from Transformers (Devlin et al., 2018).

2. Prior Work

Here, we provide a brief overview prior work in NLP that relates to deep learning and multilingual NLP. Named entity recognition and its successor, mention detection, have a vast history in NLP - a full description is beyond the scope of this paper. We will touch on the deep learning research that is directly related to the results presented here.

Collobert and Weston (2008) was the first modern approach to sequence classification, including NER, that used a convolutional neural network architecture, advancing the state-of-the-art (SotA) in English CoNLL. Lample et al. (2016) introduced – what has become the standard baseline – Bidirectional LSTM (Bi-LSTM) networks to advance the SotA NER performance on the CoNLL datasets, building 4 models, one for each language.

2018 saw the introduction of strong language-model pretrained models, first with ELMo (Peters et al., 2018), then with BERT (Devlin et al., 2018). These models excel by using large amounts of unlabeled data to train neural networks that learn the structure of the

System	En	Es	De	Nl
BERT-SL^{En} 0-shot	91.2	73.6	69.4	78.6
BERT-SL	91.2	87.5	82.7	90.6
BERT-ML	91.3	87.9	83.3	91.1

Table 1: Single and multi language F_1 on CoNLL'02, CoNLL'03 .

System	En	Ar	Zh
BERT^{En} 0-shot	87.9	10.7	65.2
BERT SL	87.9	68.7	72.9
BERT ML	88.3	69.9	74.1

Table 2: Comparison of Monolingual and Multilingual RE performance (F_1 score).

language by playing guessing games: predict the next word, predict a missing word in context, predicting the next sentence. Then, they are then used as pretrained networks to various NLP tasks, resulting in state-of-the-art results.

Vaswani et al. (2017) is the most important new development in neural network architectures for NLP which relies solely on attention mechanisms while dispensing entirely with recurrence and convolution. A particular instantiation of this architecture is BERT (Devlin et al., 2018) which trained a transformer-based architecture on large amounts of unlabeled text, with a cloze and next sentence prediction objectives, then feeding the sentence/paragraph embeddings to a linear feedforward layer, again surpassing the SotA in many tasks.

Akbik et al. (2018) extends the ELMo framework by computing Bi-LSTM sequences at character level for the entire sentence, then combines the token aligned pieces to feed into a bidirectional LSTM layer, together with the word embeddings, and obtaining SotA results on CoNLL and OntoNotes.

Multilingual Work

The resource problem or the fact that a considerable amount of time and money has to be spent in creating human labeled corpora to be used as training data for each given domain, language and NLP component has been plaguing the field since the earliest statistical models were defined and developed. As such, there has been much interesting cross-lingual induction of NLP tools, i.e. harnessing existing work in machine translation or cross-lingual dictionaries to induce NLP tools in a language without such tools from a language that does have such tools. An important early work is Yarowsky et al. (2001).

Following the development of pretrained word embeddings, interest has shifted to using these word embeddings in a multilingual setting (Ruder et al., 2017). Sil et al. (2015) trained a joint mention detection model on English and Spanish, resulting in better performance on the Spanish data. Akbik et al. (2018) did experiments by training Flair on all CoNLL'02 and '03 languages and providing one model on their github page of their system, Flair Github (2019). Xie et al. (2018) aligned monolingual embeddings from English to Spanish, German, and Dutch, and then translated the English CoNLL dataset into these languages, and built a self-attentive Bi-LSTM-CRF model using the translated languages, creating 0-shot NER systems. Pires et al. (2019) used multilingual BERT and techniques similar to our zero-shot baseline to obtain SotA numbers for zero-shot on all four CoNLL languages.

System	En	Pt	De	Es	It	Ja	Fr	Ar
SIRE	87	82	76	85	77	82	74	61
BERT^{En}	93	77	73	80	72	62	66	36
BERT ML	93	87	84	89	82	84	81	74

Table 3: Performance on the KLUE dataset, 8 languages.

Conneau et al. (2018) developed a task specifically for the multilingual setting where NLP practitioners could test knowledge transfer across languages in an unsupervised manner on a problem known as natural language inference. This is a problem where an NLP system must decide when given two sentences whether the second sentence entails the first, contradicts the first or is neither. Usually, a system would train only on labeled English sentence pairs and then be evaluated on sentence pairs from 14 difference languages that are not English such as French, Chinese, Urdu, etc. The knowledge transfer is implemented by harnessing machine translation systems that either translate non-English languages into English during decoding or translate the English training data into the target language so that a new language specific system can be trained.

3. Multilingual Named Entity Recognition

3.1. Data and Framework

We experiment on the Dutch, English, German and Spanish CoNLL data sets (Tjong Kim Sang, 2002; Sang and Meulder, 2003), the OntoNotes dataset (Weischedel et al., 2011), and the KLUE dataset, a multilingual NER dataset used in Watson NLP, and use the multilingual BERT embeddings provided by (Devlin et al., 2018). One main advantage of this method is that they create a universal vocabulary that spans the most frequent 104 languages in Wikipedia, effectively allowing us to feed many languages as input to the system.

To evaluate our hypotheses, we run two types of experiments: one in which we only train the system on the English dataset from each corpus (which is typically the largest) and then test on the other languages, and the second in which we train on all the datasets. We compare the results with the systems obtained by training on each individual language separately, as it is the common practice nowadays.

3.2. Results

Table 1 shows the results on the CoNLL dataset. The first line shows the 0-shot performance[1], which is the system trained only on English. The system that was trained on data from all languages outperforms each system trained only on its own language by an average of 0.4 F_1, which is the standard measure for NER - the hyperbolic mean of precision and recall. The 0-shot system is behind language-specific systems by 13 F_1, which is not too bad, given that the system was not exposed to the languages at all.

Table 2 shows the results obtained on the OntoNotes corpus. The interesting part here is that the languages do not share the script at all. Surprisingly, the English-trained system performs very well on Chinese, only being 7.5 F_1 behind the language-specific system, basically delivering 90% of the performance. The multilingual system is again better than the single-language systems by 0.9 F_1, including 0.4 in English, which shows that even the dataset with the largest data size can be improved using this approach.

Finally, Table 3 shows the results of running the BERT multilingual across 8 languages: English, Brazilian Portuguese, German, Spanish, Italian, Japanese, French, and Arabic. We compare here against a feature-based system developed at IBM Research - SIRE (Statistical Information and Relation Extraction) (Florian et al., 2004), which is not deep-learning based, and is representative of the best non deep learning statistical systems.

The multilingual BERT outperforms SIRE by a large margin - 10.7 F_1 on average. The English-trained BERT system is behind SIRE by 8.2 F_1 absolute (89.5% relative) and 14.4 F_1 (82.8% relative) behind the multitrained system, even though it did not have access to any of the foreign language labeled data.

3.3. Observations and Comments

These results show that if one does not have the resources to create labeled training data in a large variety of languages, they can build the data in English, and then use the trained BERT system to also have the capability of processing other languages. If one has the resources, then they can train a truly multilingual system that will perform very well across languages.

We also note that this technique is applicable to most NLP problems, not only named entity recognition - we have applied it successfully, for instance, to sentiment classification and relation extraction as well.

4. Conclusion

Multilingual pretraining in the form of multilingual BERT opens up exciting opportunities and hints at a new modus operandi for low resource languages and multilingual NLP in general. As we have shown here, one can obtain very good performance with a system that was trained only on English, and even better performance if the system is trained on multiple languages.

On three datasets, the multilingual BERT system outperformed the language-based BERT systems, and was much better than a feature-based statistical approach (SIRE). As a proxy for a single-language system, the English-trained BERT system performed at about 80-90% of the full multilingual BERT system, showing that, in cases where the resources are not there to build multiple language datasets, this is an effective approach to build a system that can tackle multiple languages at once.

We are looking forward to more research into better representation of languages that will lead to even better performance across all NLP tasks.

[1] 0-shot is used in literature to mean the system did not have any training data in that category.

References

A. Akbik, D. Blythe, and R. Vollgraf. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA, Aug. 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/C18-1139.

R. Collobert and J. Weston. A unified architecture for natural language processing: deep neural networks with multitask learning. In *Proceedings of International Conference on Machine Learning*, 2008.

A. Conneau, G. Lample, R. Rinott, A. Williams, S. R. Bowman, H. Schwenk, and V. Stoyanov. XNLI: evaluating cross-lingual sentence representations. *CoRR*, abs/1809.05053, 2018. URL http://arxiv.org/abs/1809.05053.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

R. Florian, H. Hassan, A. Ittycheriah, H. Jing, N. Kambhatla, X. Luo, N. Nicolov, and S. Roukos. A statistical model for multilingual entity detection and tracking. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 1–8, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N04-1001.

Z. R. Github. very simple framework for state-of-the-art natural language processing (nlp). https://github.com/zalandoresearch/flair, 2019.

G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer. Neural architectures for named entity recognition. *CoRR*, abs/1603.01360, 2016. URL http://arxiv.org/abs/1603.01360.

M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. *CoRR*, abs/1802.05365, 2018. URL http://arxiv.org/abs/1802.05365.

T. Pires, E. Schlinger, and D. Garrette. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, July 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P19-1493.

S. Ruder, I. Vulić, and A. Søgaard. A survey of cross-lingual word embedding models. *arXiv preprint arXiv:1706.04902*, 2017.

E. F. T. K. Sang and F. D. Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *CoRR*, cs.CL/0306050, 2003. URL http://arxiv.org/abs/cs.CL/0306050.

A. Sil, G. Dinu, and R. Florian. Proceedings of the 2015 text analysis conference, TAC 2015. NIST, 2015. URL https://tac.nist.gov/publications/2015/papers.html.

E. F. Tjong Kim Sang. Introduction to the conll-2002 shared task: Language-independent named entity recognition. In *Proceedings of the 6th Conference on Natural Language Learning - Volume 20*, COLING-02, pages 1–4, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1118853.1118877. URL https://doi.org/10.3115/1118853.1118877.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

R. Weischedel, E. Hovy, M. Marcus, M. Palmer, R. Belvin, S. Pradhan, L. Ramshaw, and N. Xue. *OntoNotes: A Large Training Corpus for Enhanced Processing*. 01 2011.

J. Xie, Z. Yang, G. Neubig, N. A. Smith, and J. Carbonell. Neural cross-lingual named entity recognition with minimal resources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 369–379, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D18-1034.

D. Yarowsky, G. Ngai, and R. Wicentowski. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*, HLT '01, pages 1–8, Stroudsburg, PA, USA, 2001. Association for Computational Linguistics. doi: 10.3115/1072133.1072187. URL https://doi.org/10.3115/1072133.1072187.

Proceedings, pages 20–22
Paris, UNESCO Headquarters, 5-6 December, 2019.

Multi-lingual Support in Connective Learning Scheme for Refining and Connecting the Open Educational Videos

Virach Sornlertlamvanich[1,2], Nannam Aksorn[2], Thatsanee Charoenporn[1]
[1]AAII, Faculty of Data Science, Musashino University
3-3-3 Ariake, Koto-ku, Tokyo 135-8181, Japan
[2]SIIT, Thammasat University
131 Moo5, Tiwanond Road, Bangkadi, Mueang, Pathumthani, 12000, Thailand
virach@musashino-u.ac.jp, nannam.aksorn@gmail.com, thatsane@musashino-u.ac.jp

Abstract

Tons of educational videos are available online. It is a big burden for learners to figure out the videos they need in the preferred time and language. Not all videos are suitable for learning according the length and presentation components. According to the Sweller's cognitive load theory, the working memory in learning process is very limited, the learner must be selective to what information from sensory memory to pay attention. In the connective learning, we effectively apply NLP approach to refine the video subtitle in archiving, translating, summarizing, classifying, and labelling the relevant keywords to create the multi-lingual learner-friendly environment.

Keywords: Connective Learning, Multi-lingual, Educational Video

Résumé

ปัจจุบันมีวิดีโอเพื่อการศึกษาที่เผยแพร่ออนไลน์มากมาย จึงเป็นการไม่สะดวกสำหรับผู้เรียนในการหาวิดีโอที่ต้องการได้ ซึ่งส่วนใหญ่ก็ต้องเลือกดูบางส่วนก่อนเพื่อให้ทราบเนื้อหา และวิดีโอส่วนใหญ่ก็เป็นภาษาอังกฤษหรือภาษาอื่นๆ ที่ผู้เรียนไม่สันทัดมากนัก จากทฤษฎีการเรียนรู้ (cognitive load theory) ของ Sweller ที่ได้กล่าวไว้ว่าในกระบวนการเรียนรู้ผู้เรียนจำเป็นต้องอาศัยหน่วยความจำชั่วคราว (working memory) ซึ่งมีพื้นที่จำกัด ดังนั้นเพื่อให้การเรียนรู้มีประสิทธิภาพสูงสุด งานวิจัยนี้ได้นำเสนอการใช้การประมวลผลภาษาธรรมชาติเพื่อช่วยในการจัดเก็บคำบรรยายประกอบ แปลคำบรรยาย ย่อความ จำแนก และสกัดคำสำคัญสำหรับการนำเสนอบทเรียนด้วยภาษาที่ต้องการและปรับแต่งให้เป็นวิดีโอที่เหมาะสมตามทฤษฎีการเรียนรู้

1. Introduction

The number of educational video titles drastically increases and covers a wide area of study. Many learners are seeking for additional learning materials to complement their understanding about lessons just learned in the classes. These educational videos are also intentionally used by the lecturers to complement the lessons taught in the classes. However, it is not easy to search and scan the tremendous files from the collections. Though they are classified by topics or the tags, it still consumes a lot of time to watch the whole bunch of the applicable videos. If learners do not watch the videos, they cannot learn from them. According to the Guo et al.'s survey, learner engagement drops off when the video length is getting longer. The median engagement time with 9-12 minute videos is about 50% and the median engagement time with 12-40 minute videos is about 20%. The maximum median engagement time for a video of any length was six minutes. Making videos longer than 6-9 minutes is therefore likely to be wasted effort. Our proposal is to archive the open license available videos and prepare them in the form to promote learner engagement and ready to learn.

2. Availability of Educational Videos

The growth of educational videos, nowadays, is leaping forward measuring in either the number of videos produced and presented on the internet or the number of users' visits. The statistics from the market survey in 2018 by Marketing Charts, as shown in Figure 1, illustrates that video viewership on Facebook and YouTube increased by up to 51.6% for business on Facebook and 23.4% on YouTube. In the educational field, there is also a significant increase of 10.2% on Facebook and 11.2% on YouTube, which is the third-highest in comparison statistics report (Erickson D. 2019).

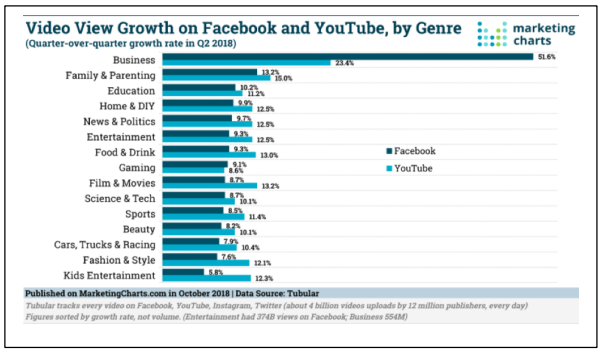


Figure 1: Video view growth on Facebook and Youtube in 2018

When there is high demand, the amount of supply that produces video for learning is higher too. It increases likely much more according to the increasing needs of online learning. There are two explicit types of online learning available on the internet. The first one is the direct learning in the system of education through the E-learning system of educational institutions such as TUXSA, the online Master degree of Thammasat University and Spillane Company (https://www.skilllane.com/academic/tuxsa?), Chula Mook of Chulalongkorn University (https: //mooc.chula.ac.th/courses). The second one is an additional learning from open courses such as Khan academy (https://www.khanacademy.org), Udemy (https://www.udemy.com), Fast AI (https://course.fast.ai), Coursera (https://www.coursera.org) as well as videos that are commonly distributed on YouTube. As the big volume of the growth of online learning, video collection services are introduced and available for learning from various

sources and for accessing easily, with the public license of YouTube, such as the Open Culture (http://www.openculture.com) where videos are categorized by academic field and accessibility. As a result, learners are able to find the exact learning video they need to learn.

3. Cognitive Load

Cognitive Load Theory, proposed by Sweller et al. (2011), suggests that memory has several components as shown in Figure 2. Sensory memory is transient, collecting information from the environment. Information from sensory memory is selected for temporary storage and processing in working memory, which has very limited capacity. This processing is a prerequisite for encoding into long-term memory, which has virtually unlimited capacity. Because working memory is very limited, the learner must be selective about what information from sensory memory to pay attention to during the learning process, an observation that has important implications for creating educational materials.

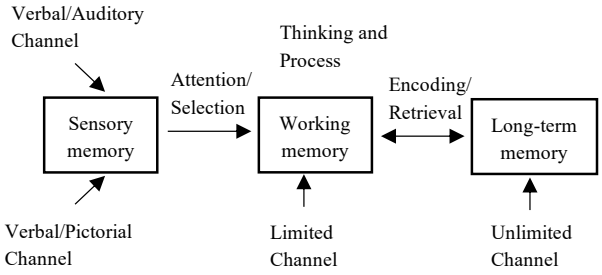


Figure 2: Components in memory based on Mayer (2003), and Mayer and Moreno (2007)

Cognitive Load Theory, proposed by Sweller et al. (2011), suggests that memory has several components as shown in Figure 2. Sensory memory is transient, collecting information from the environment. Information from sensory memory is selected for temporary storage and processing in working memory, which has very limited capacity. This processing is a prerequisite for encoding into long-term memory, which has virtually unlimited capacity. Because working memory is very limited, the learner must be selective about what information from sensory memory to pay attention to during the learning process, an observation that has important implications for creating educational materials.

However, the educational videos available on Youtube are not always appropriate for learners. According to the Cognitive Load theory, initially articulated by Sweller and colleagues (Sweller et al. 1988, 1989, 1994), suggests that memory has several components as shown in Figure 2. Because working memory is very limited, the learner must be selective about what information from sensory memory to pay attention to during the learning process, an observation that has important implications for creating educational materials. The video length is one of the significant factors that affect the learner engagement and memory as well as the structure and content of the video. This research, therefore, aims to propose a set of text processing techniques to collect, categorize and modify the videos to be suitable for human learning processing in terms of video length, video relativity and content structure according to the Cognitive Load theory.

Based on this model of memory, Cognitive Load theory suggests that any learning experience has three components. Those are,

1. **Intrinsic Cognitive Load** – This is the load imposed by the task itself. This is effectively fixed, but we should try and reduce it by 'chunking' breaking the task down into smaller parts.
2. **Extraneous Cognitive Load** – This is the environment and the way we present the information. We should try and minimize this.
3. **Germane Cognitive Load** – This is the processing that takes place comparing the new information to what we already know and encoding new learning to the long-term memory as schema. The more we know about something the lower the Germane Cognitive Load will be as thoughts and processes are automated.

The efficient learning, as shown in Figure 3, can occur when Working Memory Capacity is greater than the sum of Extraneous Cognitive Load, Germane Cognitive Load, and Intrinsic Cognitive Load (Sweller et al. 2010). Reducing extraneous load by helping novice learners with the task of determining which elements within a complex tool are important, and it can also increase germane load by emphasizing the organization of and connections within the information. Managing intrinsic load, and it can also increase germane load by emphasizing the structure of the information.

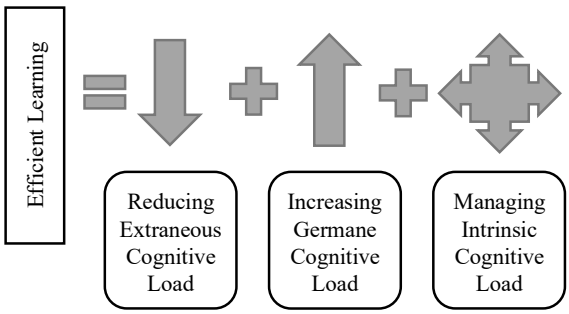


Figure 3: Working memory in learning process based on Sweller (2010)

Following the Cognitive Load theory, we propose a set of effective NLP techniques to manage the educational video resources by

1. Shortening the video length not to exceed six minutes based on the Guo et al. (2014) survey that student median engagement time for videos less than six minutes long is close to 100%.
2. Categorizing the videos for better accessibility.
3. Providing video synopsis for better structure.
4. Extracting keywords for better representation of the content.
5. Indexing the content for keyword search and scene search.
6. Summarizing the content for quick view of the contents.

Unlimite

7. Linking from one video to other related videos for total understanding.

4. Connective Learning Scheme

The system is experimentally implemented in the public cloud system. The target videos are collected with the subtitle files and archived in the cloud database. The subtitle text files are translated to any the target languages (such as Japanese, Chinese, and Thai) by Google Translate API. The resulting translation files are manipulated as source files for each language processing. Keyword extraction, summarization and video synchronization are conducted in parallel with a relating unique ID to realize the video multilingual services.

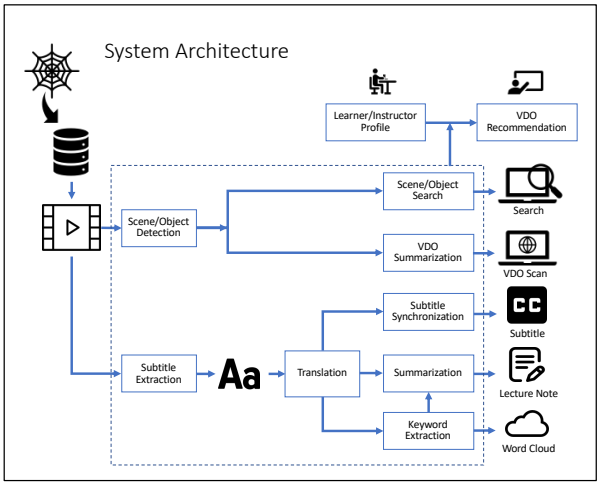


Figure 4: Connective learning system architecture

Figure 4 shows the system architecture of the proposed connective learning. The flow is started from video files crawling from the publicly open video sources available in the Internet according to their open pubic license. Video subtitles are extracted and processed in accordingly to their video contents. Text processing techniques are applied to extract the keywords, summarized and indexed.

At the same time, the video files are analyzed to detect the objects and scene representations. The preliminary experiment on video analysis is conducted to support video summarization and scene search. Finally, video recommendation based on learner view history and profile can be considered, and the instructor curriculum fulfillment function can be extended.

The system efficiently provides video playback, summary, word cloud annotated with a hyper link, scene search under the multilingual service environment. As a result, a learner can browse the summary and word cloud to understand the structure of the content before starting the video playback. A hyper link to external webpages supports the additional explanation. Scene search can direct the learner to the desired scene. The available learning videos are finally connected to realize the efficient learning environment.

Figure 5 shows one possible service of the proposed connective learning for multi-lingual learning environment.

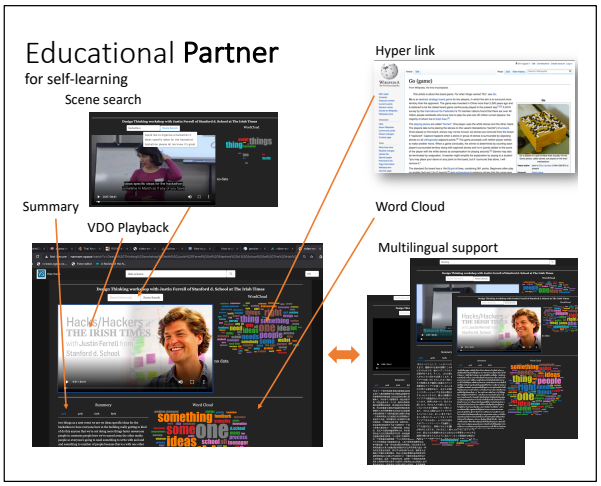


Figure 5: Connective learning system for multi-lingual service

5. Conclusion

The proposed connective learning is a result of connecting the analyzed learning video by summarizing the contents into a well-structured form of keywords, linked to external source of information. The translated contents helps reducing the inequality in education. The efficient learning can be realized by the NLP refinement based on the cognitive load theory accordingly, i.e. summarize to reduce extraneous cognitive load, scene search and external link to increase Germene cognitive load, and classify to mange intrinsic cognitive load.

6. Bibliographical References

Brame, C. J. (2015). Effective Educational Videos. https://cft.vanderbilt.edu/guides-sub-pages/effective-educational-videos/.

Erickson D. (2019). Video View Growth On Facebook & YouTube [CHART]. http://trends.e-strategyblog.com/2019/02/25/video-view-growth/30750

Guo, P. J., Kim, J., and Robin, R. (2014). How video production affects student engagement: An empirical study of MOOC videos. ACM Conference on Learning at Scale (L@S 2014)

Mayer, R. (2003). The Promising of Multimedia Learning: Using the Same Instructional Design Method across Different Media, Learning and Instruction, 12, 125-141.

Moreno, R., and Mayer, R. (2007). Interactive Multimodal Learning Environments. Educational Psychology Review, 19, 309-326.

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. Educational Psychology Review, 22, 123- 138.

Sweller, J., Ayres, P. and Kalyuga, S. (2011). Cognitive load theory. Springer. New York.

Proceedings, pages 23–26
Paris, UNESCO Headquarters, 5-6 December, 2019.

Improvement of Thai NER and the Corpus

Thatsanee Charoenporn[1], Virach Sornlertlamvanich[1,2], Kitiya Suriyachay[2]
[1]AAII, Faculty of Data Science, Musashino University
3-3-3 Ariake, Koto-ku, Tokyo 135-8181, Japan
[2]SIIT, Thammasat University
131 Moo5, Tiwanond Road, Bangkadi, Mueang, Pathumthani, 12000, Thailand
{thatsane, virach}@musashino-u.ac.jp, m5922040075@g.siit.tu.ac.th

Abstract

Thai named entity (NE) corpus is rarely found though the named entity recognition (NER) task can make a big contribution in processing the huge amount of available texts. We propose an iterative NER refinement method using BiLSTM-CNN-CRF model with word, part-of-speech, and character cluster embedding to clean up the existing NE tagged corpus due to its inconsistent and disjointed annotation. As a result, in the newly generated corpus, we obtain 639,335 NE tags, much larger than the original size of 172,232 NE tags. The generated model by the newly generated corpus also improves the NER F1-score 16.21% to mark 89.22%.

Keywords: name entity, Thai language, corpus, NE corpus

Résumé

การพัฒนาคลังข้อความภาษาไทยสำหรับการประมวลผลภาษาธรรมชาตินั้น มีประเภทและปริมาณเพิ่มมากขึ้น แต่คลังข้อความชื่อเฉพาะภาษา ไทย หรือ Thai Name Entity Corpus ยังคงมีจำนวนจำกัด แม้ว่างานวิจัยด้านการรู้จำชื่อเฉพาะ (Name Entity Recognition: NER) จะส่งผลต่อความถูกต้องของการประมวลผลข้อความเป็นอย่างมากก็ตาม งานวิจัยนี้ เสนอวิธีการปรับแต่ง NER แบบวนซ้ำ โดยใช้แบบจำลอง BiLSTM-CNN-CRF ประกอบกับ คำแวดล้อม หน้าที่ของคำ และกลุ่มอักขระข้างเคียง เพื่อปรับปรุงคลังข้อความชื่อเฉพาะภาษาไทย จากเดิม จำนวน 172,232 ชื่อ ให้มีความถูกต้อง แม่นยำ และสอดคล้องกัน ผลการวิจัยพบว่า คลังข้อความชื่อเฉพาะภาษาไทย ที่ปรับปรุงขึ้น ประกอบด้วยคำและป้ายระบุชื่อเฉพาะ (Tags) จำนวนถึง 639,335 ชื่อ ทั้งนี้ ผลการปรับปรุงคลังข้อความชื่อเฉพาะด้วยแบบจำลองที่นำเสนอนี้ สามารถกำกับชื่อเฉพาะภาษาไทยได้ถูกต้อง วัดด้วยค่า F1-score ได้ที่ 89.22 เปอร์เซ็นต์ ซึ่งให้ผลที่ดีกว่าแบบ จำลองที่สร้างด้วยคลังข้อความเดิมถึง 16.21 เปอร์เซ็นต์

1. Introduction

The performance of IE depends on many NLP preprocessing subtasks including word segmentation, POS tagging, and especially, named entity recognition (NER). NER task is to identify and classify the particular proper nouns in focus texts automatically.

Continuously, there have been researches on NER for many languages with various approaches. But NER for Thai language were still limited. There are several challenges in Thai NER. Firstly, unlike English or other European languages, there is no word boundary in Thai language. Thai words are implicitly recognized and some depend on the individual judgement. Incorrect word identification certainly affects other upper recognition than word level. As well as in NER, incorrect word segmentation will lead to false named entity recognition. Secondly, there is no capitalization in writing system to identify named entities. Even though, there are some markers in some cases identifying proper nouns like person name or institution name.

Moreover, once words are segmented and marked with named entity tags, consistency of NE tags throughout the corpus is also the important considerable issue. Since inconsistency is going to cause the failure in further processes. This paper proposes a method to clean up the existing named entity (NE) corpus and verify its consistency in creating a model for the Thai named entity recognition (NER) task. As a result, the BKD (Bangkok Data) NE corpus is newly released as an NE silver standard corpus.

2. Thai Computational Corpus

Thai is an isolate and analytic language which is spoken and written by 65 million of population inside and outside Thailand. Thai language has 44 consonant characters and 18 vowel characters with 5 tones (4 tonal symbols). Word order in sentences is the essential way to illustrate syntactic role and convey the meaning. In the writing system, Thai has no explicit word and sentence boundary as occurred in English nor the capital letter used for beginning the sentence or name entity.

Although the Thai language has a limited number of users, there is a continuous development of corpus and tools for analysis and research in the field of Thai Linguistics and Natural Language Processing. But there is still a limited number of releases for public usage. Starting from ORCHID in 1997, the first Thai POS tagged corpus which is the collaborative research between the Communications Research Laboratory and the Electrotechnical Laboratory of Japan and NECTEC of Thailand (Thatsanee et al., 1997). Now, Thai-English Parallel Corpus, Thai Speech Corpus, Thai Character Image Corpus are also publicized by NECTEC under Creative Commons Attribute 3.0 License. Thai National Corpus, initiated by Chulalongkorn University in 2006, provides the collection of current written Thai text marked up with TEI P4 standard as well as word boundaries and romanized transcription. At present, some Thai Natural Language Processing Resources including corpus, lexicons and software libraries, have been collected by Thai NLP group and can be accessed online (Kobkrit, 2019). However, the size of the corpus provided is still limited.

3. Challenges in Thai Name Entity Construction

Conforming to the Thai Name Entity (NE) Corpus, it is still limited either in number or size. THAI-NEST (THAI-Named Entities Specification and Tools) is only the open general Thai corpus with word segmentation and name entity tags. The corpus consists of over 300,000 Thai online news articles on seven major categories from twenty-one publishers which were word-segmented and tagged with seven name entity categories including person name, organization name, place name, date, time, measurement, and name (Theeramunkong et al, 2010). Figure 1 depicts NE corpus provided by THAI-NEST.

Figure 1 : THAI-NEST Corpus

From the corpus, some attempts are undertaken for NE tagged annotation. Nevertheless, some challenges can be found from the original tagged NE corpus version 0.1 (N/A). The challenges occurred during the process of NE tagging includes the correctness of word segmentation, the correctness of NE tag assigning, and the consistency of NE tag assigned along the corpus. Incorrectness of word segmentation and POS assignment sometimes have been found on the abbreviation such as "มี.ค." [March] or "อบจ." [Provincial Administration Organization]. Once they occur, as a result, some NE tags are assigned inaccurately too. Figure 2(a) and 2(c) illustrate examples of incorrectness of NE tag assignment. สอดส่อง/VACT/O in Figure 2(a), and แมนเชสเตอร์/NPRP/O and ยูไนเต็ด/NPRP/O in Figure 2(c) should be tagged as NE-PER and NE-NAM respectively instead, as shown in Figure 2(b) and (d).

พ.ต.อ./NCMN/B-PER
<space>/PUNC/I-PER
ทวี/NPRP/I-PER
<space>/PUNC/O
สอดส่อง/VACT/O

(a)

พ.ต.อ./NCMN/B-PER
<space>/PUNC/I-PER
ทวี/NPRP/I-PER
<space>/PUNC/O
สอดส่อง/NPRP/I-PER

(b)

แมนเชสเตอร์/NPRP/O
<space>/PUNC/O
ยูไนเต็ด/NPRP/O
<space>/PUNC/O
แชมป์ลีก/NCMN/B-NAM
ฤดูกาล/NCMN/O
ล่าสุด/VATT/O

(c)

แมนเชสเตอร์/NPRP/B-NAM
<space>/PUNC/O
ยูไนเต็ด/NPRP/I-NAM
<space>/PUNC/O
แชมป์ลีก/NCMN/O
ฤดูกาล/NCMN/O
ล่าสุด/VATT/O

(d)

Figure 2 : Incorrect and correct tag assignment.

บังคับ/VACT/O
ใช้/VACT/O
แล้ว/XVAE/O
ตั้งแต่/RPRE/O
เที่ยง/NCMN/O
คืน/VACT/O

(a)

จนถึง/RPRE/O
เที่ยง/NCMN/B-TIM
คืน/VACT/I-TIM
<space>/PUNC/O

(b)

Figure 3 : Inconsistency of Common Noun and NE-TIM tag assignment.

The consistency of the tagging throughout the corpus is one another challenge. From the original corpus, we found that there are some inconsistencies of tag assignment between NE and other categories, as shown in Figure 3 (a) and (b), "เที่ยง" is tagged as Common Noun (NCMN) and NE-TIM. In addition, Figure 4 (a) and (b) illustrate the examples of inconsistent and incorrect NE tag assignment between Number (DONM), Proper Name (NPRP) and NE-DAT.

วันที่/NCMN/O
23/DONM/O
<space>/PUNC/O
ถึง/RPRE/O
<space>/PUNC/O
31/NCNM/O
<space>/PUNC/O
พฤษภาคม/NPRP/O

(a)

วันที่/NCMN/B-DAT
1/DONM/I-DAT
<space>/PUNC/I-DAT
พฤษภาคม/NPRP/I-DAT

(b)

Figure 4 : Inconsistency of Number, Proper Noun and NE-DAT tag assignment.

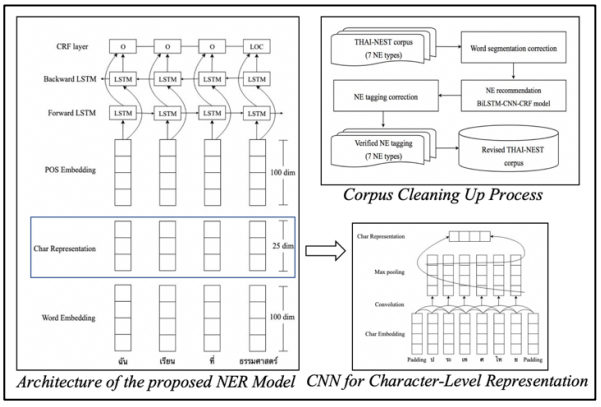


Figure 5 : Thai Name Entity Tagging Cleaning up Process

4. Thai Name Entity Tagging

We combined to create the BiLSTM-CNN-CRF model (Ma and Hovy, 2016) for predicting named entity tags. The character-level representation of each word is calculated by the CNN, as shown in Figure 5. From each embedding step, we obtain the vector representations of the words, POS tags, and character clusters. These vectors are concatenated before being fed into the Bi-LSTM layer. We apply dropout layers to both the input and output vectors of Bi-LSTM to prevent overfitting and to regularize the model. The dropout works by randomly dropping out nodes from the network during training. Finally, the output vectors of the Bi-LSTM layer are passed through the CRF layer and

decoded via the Viterbi algorithm to select the most possible sequence of the NE tag.

5. Bangkok Data Name Entity Corpus

In this part, the proposed schema and the characteristics of Bangkok Data Name Entity Corpus 2019 are described.

5.1 Corpus Mark-up Schema

There are 2 types of markers proposed to the original NE corpus, these are 1) file information markers, and 2) line number and special markers. Table 1 below displays the mark-up schema of the file information.

Mark-up	Description
%Title:	Title of the document/file
%Description:	Detail/Information of the original Text
%Number of sentence:	Total amount of sentences in the file
%Number of word:	Total amount of segmented words in the file
%Number of named entity tag:	Total of NE tags found in the file
%Date:	Date of running the NE tagger
%Creator:	Name of person who create or run the tagger
%Email:	Email address of person who create or run the tagger
%Affiliation:	Organization of the creator

Table 1: The Mark-up Schema of File Information

5.2 Number of Line and Other Special Mark-up Schema

There are 2 types of number of line marker, number of paragraph and number of sentences, as shown in Table 2 And Table 3 shows the special markers used in the corpus to convey the additional information and tags.

Mark-up	Description
#P[number]	Paragraph number of the text. The number in the bracket indicates the sequence of the paragraphs within a text.
#S[number]	Sentence number of the paragraph. The number in the bracket indicates the sequence of the sentences within a paragraph.

Table 2: The Mark-up Schema of Line Number

Mark-up	Description
\\	Line break symbol.
//	Sentence break symbol.
/[POS]	Tag marker for appropriate POS annotation of the word.
/[NE]	Tag marker for appropriate NE annotation of the word.

Table 3: The Special Mark-up Delimiters

5.3 NE Tag Annotation

According to part of speech annotation, we follow the POS tagset provided by ORCHID. NE tagset provided in the corpus consists of Date, Location, Measurement, Name, Organization, Person and Time, as displays in the Table 4. Additionally, BIO format is brought out to identify the chunk or component of NE in the sentence. “B” indicates the beginning of the chunk, and “I” presents the position of NE occurred within the chunk. While “O” is marked to identify that the word does not involve in any types of NE. Figure 6 and 7 illustrate the current Thai NE corpus in Thai and English translation.

Category	Tag	Description	Example
Date	B-DAT	Beginning of Date Name	วันที่ (Date)
	I-DAT	Inside of Date Name	1 มกราคม (January 1)
Location	B-LOC	Beginning of Location Name	จังหวัด (province)
	I-LOC	Inside of Location Name	ปทุมธานี (Pathumthani)
Measure-ment	B-MEA	Beginning of Measurement Name	สาม (Three)
	I-MEA	Inside of Measurement Name	คัน (Car)
Name	B-NAM	Beginning of Proper Name, except Location, Person and Organization Name	ศึก (League)
	I-NAM	Inside of Proper Name	ลาลีกา (La Liga)
Organiza-tion	B-ORG	Beginning of Organization Name	บริษัท (Corp.)
	I-ORG	Inside of Organization Name	เสริมสุข (Sermsuk)
Person	B-PER	Beginning of Person Name	นาง (Mrs.)
	I-PER	Inside of Person Name	สุเทพ เทือกสุบรรณ (Suthep Thaugsuban)
Time	B-TIM	Beginning of Time	เก้า (Ten)
	I-TIM	Inside of Time	โมง (O’clock)
Other	O	Does not belong any types	

Table 4: The Thai Name Entity Tagset

6. Conclusion

We adopted a collection for NE corpus originally prepared by THAI-NEST and undertook POS and NE tagged by N/A, by verifying the annotation consistency and iteratively re-annotated it with the created model. We extensively conducted the cross annotation among the seven NE tagged files of THAI-NEST to increase the number of NE tags and to prepare for additional NE tag context capturing in NER model development. The newly generated corpus consists of 639,335 NE tags. The revised NE tagged corpus with the BiLSTM-CNN-CRF model with word, part-of-speech and character embedding approach improves the NER F1-score 16.21% to mark 89.22%. Our next step is to undertake the NE tagging to ORDHID and distribute for research purpose.

```
%Title: BKD-7 corpus
%Description: This corpus based on the original THAI-NEST corpus and combined seven types of entities: DATe, LOCation, MEAsurement, NAMe, ORGanization, PERson, TIMe
%Number of sentence: 419
%Number of word: 53,319
%Number of named entity tag: 11,891
%Date: July 31, 2019
%Creator: Kitiya Suriyachay and Virach Sornlertlamvanich
%Email: m5922040075@g.siit.tu.ac.th and virach@siit.tu.ac.th
%Affiliation: Sirindhorn International Institute of Technology, Thammasat University

#S14

เมื่อเวลา 13.00 น. วันที่ 6 พฤษภาคม พ.ต.ท.ไชยศ มุกดาหาญ รอง ผกก.ป.สภ.นครชัยศรี จ.นครปฐม ได้รับแจ้งมีอุบัติเหตุรถบรรทุกสิบล้อ\\ชนกับรถยนต์ มีผู้ได้รับบาดเจ็บ 2 ราย เหตุเกิดที่บริเวณ ถ.เพชรเกษมขาเข้า หมู่ 5 ต.ศีรษะทอง อ.นครชัยศรี จ.นครปฐม …//

เมื่อ/JSBR/O
เวลา/NCMN/O
<space>/PUNC/O
13.00/DCNM/B-TIM
<space>/PUNC/I-TIM
น./CMTR/I-TIM
<space>/PUNC/O
วัน/NCMN/B-DAT
ที่<space>6/DONM/I-DAT
<space>/PUNC/I-DAT
พฤษภาคม/NCMN/I-DAT
<space>/PUNC/O
พ.ต.ท./NTTL/B-PER
ไชยศ/NPRP/I-PER
<space>/PUNC/I-PER
มุกดาหาญ/NPRP/I-PER
…
ผู้/PPRS/O
ได้รับ/VSTA/O
บาดเจ็บ/VSTA/O
<space>/PUNC/O
2/DCNM/B-MEA
<space>/PUNC/I-MEA
ราย/CNIT/I-MEA
<space>/PUNC/O
เหตุ/NCMN/O
เกิด/VSTA/O
ที่/RPRE/O
บริเวณ/NCMN/O
…
//
```

Figure 6: Thai Name Entity Corpus (Thai)

```
%Title: BKD-7 corpus
%Description: This corpus based on the original THAI-NEST corpus and combined seven types of entities: DATe, LOCation, MEAsurement, NAMe, ORGanization, PERson, TIMe
%Number of sentence: 419
%Number of word: 53,319
%Number of named entity tag: 11,891
%Date: July 31, 2019
%Creator: Kitiya Suriyachay and Virach Sornlertlamvanich
%Email: m5922040075@g.siit.tu.ac.th and virach@siit.tu.ac.th
%Affiliation: Sirindhorn International Institute of Technology, Thammasat University

#S14

At 13.00 hrs., on May 6, Pol. Col. Chaiyos Mukdahan, Deputy Director of Nakhon Chai Si Police Station, Nakhon Pathom, was informed that a ten-wheel truck accident collided with a car. There are 2 people injured in the accident. The accident occurred at the Inbound of Petchkasem Rd., Village No. 5, Sisa Thong Subdistrict, Nakhon Chai Si District, Nakhon Pathom Province. //

at/JSBR/O
time/NCMN/O
<space>/PUNC/O
13.00/DCNM/B-TIM
<space>/PUNC/I-TIM
o'clock/CMTR/I-TIM
<space>/PUNC/O
day/NCMN/B-DAT
<space>sixth/DONM/I-DAT
<space>/PUNC/I-DAT
May/NCMN/I-DAT
<space>/PUNC/O
Pol.Col./NTTL/B-PER
Chaiyos/NPRP/I-PER
<space>/PUNC/I-PER
Mukdahan/NPRP/I-PER
…
man/PPRS/O
was/VSTA/O
injured/VSTA/O
<space>/PUNC/O
2/DCNM/B-MEA
<space>/PUNC/I-MEA
person/CNIT/I-MEA
<space>/PUNC/O
accident/NCMN/O
occur/VSTA/O
at/RPRE/O
area/NCMN/O
…
//
```

Figure 7: Thai Name Entity Corpus (English Translation)

7. Bibliographical References

Kobkrit (2019). Thai NLP Resource. Retrieved from https://github.com/kobkrit/nlp_thai_resources.

Thatsanee Charoenporn, Virach Sornlertlamvanich and Hitoshi IsaharaCastor, Building A Large Thai Text Corpus---Part-Of-Speech Tagged Corpus: ORCHID. Proceedings of the Natural Language Processing Pacific Rim Symposium, 1997.

Theeramunkong, T., Boriboon, M., Haruechaiyasak, C., Kittiphattanabawon, N., Kosawat, K., Onsuwan, C., Siriwat, I., Suwanapong, T., and Tongtep, N. (2010). THAI-NEST: A framework for Thai named entity tagging specification and tools. *Proceedings of the 2nd International Conference on Corpus Linguistics (CILC10), Spain.*

X, Ma., and E, Hovy. (2016). "End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF," *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Proceedings, pages 27–30
Paris, UNESCO Headquarters, 5-6 December, 2019.

Toward Narrative-Based Conversational Interfaces

Ethel Chua Joy Ong
Center for Language Technologies, De La Salle University
2401 Taft Avenue, Malate, Manila, 1004 Philippines
ethel.ong@dlsu.edu.ph

Abstract

Advances in language technology research led to the rise in popularity of conversational interfaces that allow human users to carry on natural-like conversations with the software agent during task performance. Children comprise a sector of society who can benefit from interacting with these agents, particularly to enhance their language acquisition, literacy skills and emotional development. Stories dominate the everyday conversations of children. They learn about themselves and their environment through sharing of stories. In this paper, we describe our conversational storytelling agent that can support children in constructing a story, reflecting on story events, and expressing their emotions.

Keywords: conversational interfaces, conversational agents, collaborative storytelling, joint story reading, dialogue

1. Introduction

The rising popularity of conversational interfaces, notably voice assistants such as Apple's Siri, Amazon's Alexa, Google Assistant and Microsoft Cortana, is geared toward setting up a more natural interface wherein the human user and the computer are able to interact meaningfully when carrying out certain tasks. Despite the achievement of these technologies in terms of the number of available *skills* and *actions* (Dale, 2019), their capabilities are still limited to specific goal-oriented functions, such as searching for information to answer user queries, recommending a route, reminding users about upcoming events, playing the user's favorite music, and scheduling appointments for the user.

Fay (2014) believed that these voice assistants cannot yet engage their human users in everyday conversations akin to natural human discourse. A key missing ingredient, he noted, is the conversational agent's capacity to understand and generate stories, which is inherently embedded in the human thought process. Everyday human conversations are marked by features of storytelling, becoming free-flowing exchange of information on one's life events and experiences. Narrative intelligence affords people the ability to organize a given series of events into stories (Blair and Meyer, 1997).

As computers become more ubiquitously involved in our day-to-day activities, software agents used in conversational interfaces should embody some form of narrative intelligence to enrich their conversations with human users. Storytelling strategies can be used to elicit further details regarding a user's needs, and to analyze user requests in the context of his/her environment.

Recent advances in language technologies can support the development of conversational agents exhibiting storytelling abilities, with applications in collaborative child-agent storytelling and joint story reading, to encourage children to express their ideas while developing their linguistic and literacy skills. The agent can function as a facilitator, a tutor, or a learning peer, thus providing opportunities for itself and the human user to augment each other's storytelling abilities. In this paper, we describe our conversational storytelling agents, discuss the challenges we encountered during their development, and recommend further work to extend their capabilities.

2. Related Work

Conversational agents are virtual agents designed to mimic human capabilities and attributes, and to collaborate with human users to facilitate thinking and decision-making during the performance of specific tasks. They utilize various means of communication, from written text to voice inputs, in order to "engage in dialogues, and negotiate and coordinate the transfer of information" (Coen, 1995).

2.1 Roles of Conversational Agents

Conversational agents are currently predominantly found in commercial applications as personal voice assistants and service-performing agents. Designed primarily to provide customer support, they can assist in product search, answer customer queries, and perform simple well-defined tasks requested by the user.

Since conversational agents' roots can be traced to intelligent software agents, they can also take on various roles depending on the types of interaction expected from them, as shown in previous research. These include counselor, critic, facilitator, tutor, and peer.

As a counselor, the agent promotes certain behaviors and practices, such as in healthcare. The work of Bickmore et al. (2013) describes a health counselor agent that uses conversations to promote healthy behaviors among its users, specifically performing physical activities, and consuming fruits and vegetables. A study by Fitzpatrick et al. (2019), on the other hand, investigates how Woebot's cognitive behavior therapy skills can help users exhibiting symptoms of anxiety and depression.

As a critic, the agent is assigned to "look over the shoulder" of the users as they perform their tasks, and to offer appropriate advice as the need arises (Terveen, 1995). Critics simulate human problem-solving strategies to identify and inform users of potential concerns during task performance and offer alternative perspectives. The critic must be able to present "a reasoned opinion about a user's product or action", and must "recognize and communicate issues concerning a product" (Fischer et al., 1991a). The critic should also be able to engage the user in problem-solving tasks to resolve issues, and to provide alternative solutions by presenting their advantages and limitations for consideration (Fischer et al., 1991b).

In learning environments, conversational pedagogical agents can vary their level of intelligence and the manner by which they communicate with the learners. These abilities allow them to function not only as experts (Baylor and Kim, 2005), tutors (Graesser et al., 2005) and mentors (Zakharov et al., 2007), but also as learning companions (Cassell et al., 2005) and teachable agents (Zhao et al., 2012). In the latter two cases, the pedagogical agents serve as peers that exhibit collaborative behavior in learning activities through conversations and coaching (Ryokai, Vaucelle and Cassell, 2002), in order to promote a two-way exchange of knowledge.

Because people are inherently social beings, research has also been devoted to developing chatbots that can be interact socially and can engage in empathic conversations. One example is Microsoft XiaoIce, whose main goal is to build a long-term emotional connection with its users. Studies conducted by Zhou et al. (2018) showed that with continued use, users would eventually consider XiaoIce as a friend, frequently engaging in daily conversations with the chatbot by sharing their hobbies and interests.

2.2 Dialogue

A dialogue is defined as a two-way communication activity between two entities, usually with the intent of seeking information, negotiating, persuading, and deliberating in highly contextual tasks and discussions. To facilitate human-agent interaction, both entities must express their intentions through a series of dialogue exchange.

Dialogue exchange is characterized in terms of communicative goals or speech acts. Common speech acts include: inquiry or direct question to solicit additional information; informing to respond to a request for information; elaboration to provide additional definitions or descriptions; justification to explain actions; motivation to persuade someone to carry out an action; exemplification to demonstrate how the task can be carried out, and repair to resolve misunderstanding.

During storytelling, the dialogue exchange between the conversational agent and the user may center around the story itself and its elements, such as characters, setting and plot events, or everyday events that are interesting for the user and that may or may not be directly related to the story. The participating entities can retell story events, describe character attributes, justify character actions and motivations, imagine possible outcomes resulting from character actions or variations of existing stories, and reflect on the relevance of story events to one's daily experiences. This "ability to imagine, solve problems and make decisions", according to Fay (2014), is what makes "story generation play a vital role in intelligence".

3. Conversational Storytelling Agents

Collaborative storytelling entails a two-way exchange of ideas, feedback and suggestions regarding a story. Co-authoring through taking turns to generate story content can also occur during the collaboration. Written text or voice-based interfaces can be used as the medium of communication, both with their respective advantages and constraints as reported in (Ong et al., 2019).

3.1 Orsen, the Storytelling Peer

Orsen is a collaborative storytelling peer with input understanding and text generation abilities designed to help children create their own stories. It uses Google Firebase as the text-based interface, and Google Home and Amazon Alexa as voice-based interfaces. It utilizes a number of dialogue moves to formulate an appropriate response for a given user input. Pumps, such as "*I see, what happens next?*" and "*Tell me more about <character>.",* are used to elicit additional details regarding a story element. Hints, e.g., "*Then Salie went to market*.", allow Orsen to suggest content in order to help a child who is stuck and does not know how to proceed with his/her story.

A collection of commonsense concepts and their relations extracted from children's stories, described in (Ong et al., 2018), is used by Orsen to process user input and to generate a response. This gives children the perception that they are collaborating with an intelligent storytelling peer who understands concepts and their interrelationships normally present in the physical human world. In this way, Orsen can augment the human writer's limited knowledge by offering prompts that could lead to writing ideas.

Because storytelling is a mutual learning process for the entities involved, Orsen also behaves as a teachable agent. It leverages on the principle of learning by teaching (Zhao et al., 2012) when it encounters an unfamiliar concept, i.e., a concept that is not in its knowledge repository. In this situation, Orsen switches to a learner role and generates inquisitive responses such as "*I want to hear more about bandages*." This knowledge acquisition strategy helps in expanding the agent's knowledge base, but may also lead to learning new assertions that are not necessarily true or acceptable to all users. To address this, Orsen makes a suggestion in subsequent dialogue turns to verify the validity of its acquired knowledge. For example, to verify that a "*swing*" can be found in the "*park*", the agent can ask "*Did she see a swing?*" in response to the child's input text "*The princess walks to the park.*"

Analysis of conversation logs showed that stories shared by children vary depending on their interests and reading habits. Traces of everyday experiences are evident in some stories, expressed through character actions and events. Retelling of popular fables and fairy tales manifest in others. The length of the stories are constrained by the child's linguistic skills and personality. Some children willingly share their stories with Orsen, while others struggle with writer's block. The mode of communication also affected the interaction. Text-based interfaces require correct grammar and spelling, while voice-based interfaces require appropriate pronunciation and accent; otherwise, communication breakdowns would emerge due to Orsen's inability to understand the user's input (Ong et al., 2019).

Most language technologies are based on English, which is also one of the main languages comprising the bilingual nature of most everyday conversations in the Philippines. While Orsen is currently designed to process and generate English text, the presence of common Filipino words, e.g., *tita* (aunt) and *lola* (grandmother), has been detected in a handful of the stories.

Design challenges include establishing the size of the knowledge base to make the agent act intelligently; determining the domain of the assertions comprising this

knowledge base to adequately cover the wide variety of topics that children may share; designating the mode of communication to be used in interacting with children (text or voice); and handling communication failure. Recent advances in language technologies may be able to address speech synthesis issues, produce agent responses free from grammar errors, and correct parse user inputs to properly identify story elements regardless of misspellings and mixed language insertions. A balance must be achieved between a highly intelligent agent capable of following any story topic conceived by the children, and an agent that simply utters "That's interesting, please tell me more!" (to hide its inability to generate a relevant story content), but still leaves enough space for children to take center stage in the story creation process.

3.2 Towards Emotional Intelligence with Eren

Although stories created with Orsen ultimately remain fictional in nature, children may have drawn from personal experiences to craft their narrative. It was observed that some children shared stories about real-life events that affected them emotionally. On the basis that children can be asked to talk about their emotions freely through storytelling (Denham et al., 1996), this then paves an opportunity for developing emotional intelligence (EI). EI involves recognition of one's emotions, sensitivity to other people's emotions, and control of raging emotions.

Built on top of Orsen, Eren is a conversational agent that treats stories shared by children as accounts of their personal experiences. Based on the extracted characters, objects and events from the child's input, Eren identifies the associated emotion using the OCC Model (Shaikh et al., 2009); validates this with the user; then prompts the user to reflect on his/her actions. Eren adapts Orsen's dialogue moves using the emotion coaching model of Gottman et al. (1996) to help children understand and resolve their emotions during storytelling. In contrast to Orsen's free-flowing dialogue controlled by the child, Eren follows a set sequence of phases in its dialogue. This is necessary so that the conversation flows from emotion recognition, to cause identification, to action reflection ("*Do you think what <character> did was right*?"), and lastly to post-evaluation ("*What did you feel after telling me this*?").

Combining AI with an emotion-aware agent poses certain design challenges. Eren's emotion recognition ability is not perfect; at times, it may fail to identify the emotion from the child's input. Thus, it requires a confirmation dialogue ("*You seem <emotion>, is that right?*") to validate the detected emotion proceeding to the next phase of the conversation. This, however, has led to situations wherein the child indicated that he/she does not feel anything.

Certain pumps from Orsen – particularly those that ask for character and object descriptions, such as "*How old is <character>?*" and "*How big is <object>?*" – were found to distract the child from focusing on the emotional impact of actions and events taking place in his/her story. Because they do not support the agent's listener role, these types of pumps were subsequently removed from Eren.

Results from end user feedback showed that most children acknowledged Eren as a human ("*I feel like talking to a real human*"), a friend ("*Like a friend who won't judge you*") or a therapist ("*therapist vibe*"). Children also identified the preservation of privacy as an important factor in deciding to share their personal stories with a conversational agent ("*(Eren) won't share my story with others*" and "*Not a person so it won't judge me*"). This affirms the findings reported by Fryer and Carpenter (2006) that people tend to be more willing to disclose emotional stories to a conversational agent since it does not lose patience and does not judge.

3.3 Charm, a Chatbot for Joint Story Reading

Soriano and Ong (2016) explored how emotional changes could be elicited from the learner through the use of text-based conversation involving system-generated questions and user responses. By disrupting and shifting the learner's negative affect to a more positive perception toward learning, the software agent was envisioned to indirectly encourage the learner to complete the learning task at hand. A key design issue revealed in this work, however, was that the virtual peer evokes a conversation only when the learner is taking the reading comprehension test, which is the last part of the learning activity. But end user testing showed that the learners would have preferred an earlier intervention, specifically during the reading task itself. Such results motivated the development of Charm.

Charm is a conversational agent designed to mimic the dialogue exchange that takes place during story reading between an adult and a child. It aims to address three levels of concerns – comprehension, relevance and engagement – which affect the child's motivation to finish a given reading material. These concerns in turn are supported through three types of dialogues: cognitive dialogue to facilitate comprehension of the text; reflective dialogue to relate story events to the reader's personal life; and elaborative dialogue to promote engagement with the story by focusing on the key characters and events (Chan and Ong, 2018).

While the user has control over the conversation flow, Charm formulates its response following the mental model and proposition theories of Gunning (1996). Except for *who* and *where* questions, the agent does not give direct answers to other types of questions posed by the user. Instead, Charm rephrases the questions and throws them back to the user, encouraging the latter to seek answers to his/her own question by recalling story elements. Question formulation uses the question circles strategy (Tofade et al., 2013) to construct subject matter questions as well as external reality questions (Chan and Ong, 2018).

Mixed feedback was received from participants (Chan and Ong, 2018). On a positive note, Charm was found to be "*easy to talk to*" despite the need to use the English language which reduces the pleasantness when compared to conversing with a classmate or a friend in mixed languages. The uniqueness of the experience also piqued the interest of the participants, which they said is very different from having a face-to-face conversation. However, the participants' inherent lack of eagerness to read the story cause major usability issues, in particular, as the agent would keep asking the users to think of the answer on their own; some participants started feeling disappointed and even frustrated with the agent.

As an intelligent agent, Charm's library is currently limited to only one reading material. It has a manually-built computational model of the major scenes and characters involved in the story. It then uses this model to retrieve answers to *who* and *where* questions, and to formulate

what, *why*, *how* and *do-you-think* questions about character actions and events. Advances in language technology could enable Charm to automatically extract relevant elements from a given input story text.

4. Conclusion and Future Work

Stories abound in everyday human conversations. We explored the potential applications of voice assistants and chatbot interfaces in collaborative storytelling to provide interactive spaces for children to create their own stories, share personal experiences, and introspect/reflect on emotions. We encountered design challenges that affected how children perceived the usability of these agents. Issues concerning effective speech synthesis particularly for bilingual and multilingual users, usage of correct grammar structures in composing written and oral text, and the adequacy and extent of the capabilities of existing language technologies need to be addressed in future work to enrich the child-agent interaction.

5. Acknowledgements

This research is funded through grants from the Department of Science and Technology – Philippine Council for Industry, Energy and Emerging Technology Research and Development (DOST-PCIEERD), and from De La Salle University.

6. Bibliographical References

Baylor, A.L. and Kim, Y. (2005). Simulating instructional roles through pedagogical agents. *Intl. J. of Artificial Intelligence in Education*, 15(2):95-115.

Bickmore, T., Schulman, D. and Sidner, C. (2013). Automated interventions for multiple health behaviors using conversational agents. *Patient Education and Counseling*, 92(2):142-148, Elsevier Ireland.

Blair, D. and Meyer, T. (1997). Tools for an interactive virtual cinema. In Trappl, R. and Petta, P. (eds.) *Creating Personalities for Synthetic Actors: Towards Autonomous Personality Agents*, pages 83-91. New York: Springer.

Cassell, J., Tartaro, A., Rankin, Y., Oza, V. and Tse, C. (2005). Virtual peers for literacy learning. *Educational Technology special issue on Pedagogical Agents*.

Chan, L. and Ong, E. (2018). Engaging children in conversations during story reading. In *Proc. of the 26th ICCE Workshop on Innovative Technologies for Enhancing Interactions and Learning Motivation*.

Coen, M.H. (1995). SodaBot: A software agent construction system, MIT AI Lab, USA.

Dale, R. (2020). Voice assistance in 2019. *Natural Language Engineering*, 26, pages 129-136. Cambridge University Press.

Denham, S., Mason, G. and Auerbach, S. (1996). Mother-child dialogue about emotions and preschoolers' emotional competence. *Genetic Psychology Mono-graphs*, 121(3):311-337.

Fay, M.P. (2014). *Driving story generation with learnable character models*, PhD Dissertation, Massachusetts Institute of Technology.

Fischer, G., Lemke, A.C., Mastaglio, T. and Morch, A.I. (1991a). The role of critiquing in cooperative problem solving. *ACM Transactions on Information Systems*, 9(3):123-151.

Fischer, G., Lemke, A.C., McCall, R. and Morch, A.I. (1991b). Making argumentation serve design. *Human-Computer Interaction*, 6(3-4):393-419.

Fitzpatrick, K.K., Darcy, A. and Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR Mental Health*, 4(2).

Fryer, L. and Carpenter, R. (2006). Bots as Language Learning Tools. *Language Learning and Technology*, 10(3):8-14.

Gottman, J.M., Katz, L. F. and Hooven, C. (1996). Parental meta-emotion philosophy and the emotional life of families: Theoretical models and preliminary data. *J. of Family Psychology*, 10(3):243-268.

Graesser, A. C., Olney, A., Haynes, B. C. and Chipman, P. (2005). AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education*, 48(4):612-618.

Gunning, 1996 Charm Thomas Gunning. 1996. *Creating Reading Instruction for All Children,* 2nd ed. Allyn & Bacon, Des Moines, IA.

Ong, D.T., Gilig, L.K., De Jesus, C.R., Alburo, J.B. and Ong, E. (2018). Building a commonsense knowledge base for a collaborative storytelling agent. In Yoshida K., Lee M. (eds.) *Knowledge Management and Acquisition for Intelligent Systems*, LNAI 11016, pages 1-15.

Ong, E., Alburo, J.B., De Jesus, C.R., Gilig, L.K. and Ong, D.T. (2019). Challenges posed by voice interface to child-agent collaborative storytelling. In *Proc. of the Oriental COCOSDA 2019*, IEEE.

Ryokai, K., Vaucelle, C. and Cassell, J. (2002). Literacy learning by storytelling with a virtual peer. In *Proc. of the Conference on Computer Support for Collaborative Learning: Foundations for a CSCL Community*, pages 352-360, Colorado.

Shaikh, M., Prendinger, H. and Ishizuka, M. (2009). A linguistic interpretation of the OCC emotion model for affect sensing from text. *Affective Information Processing*, pages 45-73, Springer-Verlag London.

Soriano, Z. and Ong, E. (2016). A conversational agent to shift students' affect state. In M. Baldoni et al. (eds.), IWEC 2014 / IWEC 2015 / CMNA 2015, *Principles and Practice of Multi-Agent Systems*, LNAI 9935, pages 1-12, Springer International Publishing, Switzerland.

Terveen, L. G. (1995). An overview of human-computer collaboration. *Knowledge-Based Systems*, 8(2-3):67-81.

Tofade, T., Elsner J. and Haines, S. (2013). Best practice strategies for effective use of questions as a teaching tool. *American J. of Pharmaceutical Education,* 77(7):155.

Zakharov, K., Mitrovic, A. and Johnston, L. (2007). Pedagogical agents trying on a caring mentor role. In *Proc. of the 2007 Conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work*, pages 59–66. IOS Press.

Zhao, G., Borjigin, A., and Shen, Z. (2012). Learning-by-teaching: Designing teachable agents with intrinsic motivation. *J of Educational Technology & Society* 15(4):62–74.

Zhou, L., Gao, J., Li, D., and Shum, H.Y. (2019). The design and implementation of XiaoIce, an empathetic social chatbot. *Computational Linguistics*, 1(1):1-62.

Proceedings, pages 31–35
Paris, UNESCO Headquarters, 5-6 December, 2019.

SukatWika: An Analysis Software for Linguistic Properties of Texts

Kathrina Lorraine Lucasan[1], **Angelina Aquino**[2], **Francis Paolo Santelices**[3], **Dina Ocampo**[4]
[1,4]Center for Integrative and Development Studies - Education Research Program
[2,3]Electrical and Electronics Engineering Institute
University of the Philippines, Diliman, Quezon City, Philippines
{kmlucasan, dina.ocampo}@up.edu.ph, {angelina.aquino, francis.santelices}@eee.upd.edu.ph

Abstract

There is a lack of understanding on the qualities of texts that children can read, especially in Philippine languages. Text quality should be informed by an analysis of text difficulty, which can be measured by the linguistic properties of text such as word density, concept load, and phonological weight. The SukatWika analysis tool was developed to automate the extraction of this information for texts written in Filipino, English, Sinugbuanong Binisaya, and Ilokano languages. The results obtained from this software can be used as an aid in the creation of instructional materials which will support reading development among learners.

Keywords: text analysis software, linguistic properties, Philippine languages, literacy assessment

Buod

Mayroong kakulangan sa pag-unawa sa kalidad ng tekstong nababasa ng mga bata, lalo na sa mga wika sa Pilipinas. Dapat nakahango ang kalidad ng teksto sa pagsusuri ng antas nito, na maaaring sukatin mula sa mga katangiang panlinggwistika ng teksto tulad ng haba at dami ng salita, *concept load*, at sa mga uri ng tunog na nakapaloob sa mga salita. Ang SukatWika ay binuo upang mabilisang makuha ang mga impormasyong panlinggwistika ng tekstong nasa wikang Filipino, Ingles, Sinugbuanong Binisaya, at Ilokano. Ang impormasyong maibibigay ng SukatWika ay makatutulong sa paggawa ng mga kagamitang panturo upang malinang ang kakahayan ng mga mag-aaral sa pagbasa.

1. Introduction

Mother Tongue-Based Multilingual Education (MTB-MLE) is an education program for children wherein they learn to read and write first in their mother tongue (MT) and then use their MT as they learn to understand, speak, read, and write in other languages (UNESCO, 2018). MTB-MLE programs may be for maintenance, for transition, or for enrichment (Lin and Man, 2009). In the Philippines, the MTB-MLE program is implemented from Kindergarten to Grade 3. The program is transitional as the goal is to provide learners with the support they need to gradually and effectively move from MT instruction to mostly Filipino and/or English instruction (Department of Education, 2019). Learning literacy in the mother tongue serves as the foundation for learning literacy in other languages (Cummins, 2003).

The Department of Education released several policies for its implementation, with one released in 2009 to institutionalize the program for Kindergarten to Grade 3, and two others in 2012 and 2013 containing the guidelines and the list of official languages, including Filipino, English, Sinugbuanong Binisaya, and Ilokano. Section 5 of Republic Act (RA) 10533 or the Enhanced Basic Education Act of 2013 stipulated the features of the K to 12 curriculum and mandated that it adhere to the principles and framework of MTB-MLE. The most recent policy was released in 2019, which articulated provisions further, including a guide for possible classroom scenarios.

Many studies discuss the advantages of the use of the MT in schools including increased classroom participation, positive affect, and increased self-esteem (UNESCO, 2004), flexibility with learning strategies (Dahm and De Angelis, 2018), faster learning of a second language (Monje et al., 2019), and increased academic achievement (Nguyen, 2017).

Because the MTB-MLE program has only been recently implemented in the Philippines, many aspects for improvement have been observed (Monje et al., 2019). Cordero (2019) for example, includes the availability of assessment tools and resources among topics that need more research.

1.1. Multi-Literacy Assessments for Filipino Learners

Multi-Literacy Assessments for Filipino Learners is a battery of assessment tools which will measure learners' skills in various literacy domains. It is part of a larger research agenda investigating lifespan literacy development of Filipinos. For its first phase, it aims to develop assessment tools for early literacy in Filipino and English and the mother tongues Sinugbuanong Binisaya and Ilokano. The literacy skills to be assessed include oral and written language development, alphabet knowledge, spelling, decoding, and listening and reading comprehension.

Among the initial steps of development was to conduct an inventory of existing assessment tools made by researchers of the university and secure their permission to adopt/adapt their tool. A priority step in the assessment package development process was to find a way to ensure that test items in the tools were drawn from texts and books that typical Kindergarten to Grade 3 elementary students in the Philippines ordinarily encounter.

2. SukatWika Analysis Tool

In line with the goals of this project, the SukatWika (Filipino: lit. "Measure Language") program was developed to

create a corpus of children's language and texts and to facilitate the analysis of linguistic properties in Philippine texts. Specifically, it automates the counting of various metrics at the sub-word, word, and sentence level which are typically used to assess the surface difficulty levels of reading and instructional materials.

2.1. Functionality

As of Version 1.0, SukatWika is capable of analyzing texts written in four Philippine languages: Filipino, English, Sinugbuanong Binisaya, and Ilokano. Given a text document in these languages, the program then provides the lengths and frequencies of lexical and grammatical units, as well as an interface for searching lexical units within the text.

Paragraph length counter. This displays the total number of paragraphs in the text, and enumerates the frequency of paragraph lengths, answering the question "How many paragraphs contain n sentences?"

Sentence length counter. This displays the total number of sentences in the text, and enumerates the frequency of sentence lengths, answering the question, "How many sentences contain n words?"

Phoneme counter. This displays the total number of phonemes in the text, tabulates the frequency of appearance of individual phonemes, and enumerates the frequency of word lengths by phoneme, answering the question "How many words contain n phonemes?" It also gives a list of unique words in the text, ordered by the number of phonemes in each word.

Word frequency counter. This displays the total number of words in the text, tabulates the frequency of appearance of unique words in the text, and sorts the words alphabetically and by frequency.

Word length counter. This displays the total number of syllables in the text, and enumerates the frequency of word lengths, answering the question "How many words contain n syllables?" It also gives a list of unique words in the text, ordered by the number of syllables in each word.

Word searcher. This allows the user to input a string of characters, and gives a list of words containing the string. It also provides options to filter the words displayed by the number of syllables contained in the word, as well as the position of the string in the word (i.e. start, middle, or end of the word).

2.2. Orthography

For the Philippine langauges supported by the program, the rules for syllabication and phonemic transcription of individual words were based on official orthographies for Filipino (Almario (Ed.), 2014), Sinugbuanong Binisaya (Akademiyang Bisaya, 2011), and Ilokano (Komisyon sa Wikang Filipino, 2012). For English, phonemic transcriptions were extracted from the CMU Pronouncing Dictionary (Carnegie Mellon University, 2014), while syllable counts were performed by simply counting the number of vowels in the phonemic transcription, since each syllable in an English word is known to contain only one sonant or vowel sound (Malone, 1957).

Tokenization rules were identical for all four languages: words were tokenized based on whitespace, while sentences and paragraphs were tokenized based on end-of-sentence and newline characters. The parsing rules for each type of analysis were then encoded as Python functions and used in the succeeding scripts to produce the necessary outputs.

2.3. User Interface

In order for the functions to be user-friendly and to be able to visualize the results, a graphical user interface was created based on PyQt5. PyQt5 is a comprehensive set of Python bindings for Qt5. Qt is set of cross-platform C++ libraries that implement high-level APIs (Application Programming Interfaces) for accessing many aspects of modern desktop and mobile systems. Shown in Figure 1 is the starting window wherein no text has been analyzed yet.

Figure 1: Starting window of the SukatWika user interface

Each of the tabs that can be seen in the lower part of the user interface displays each of the features mentioned in the previous section. For example, the test results of the word length counter can be seen in Figure 2. Also seen below the tabs is the status bar which displays the path to the current analyzed file, the language selected, and instructions to export the results. The full SukatWika analysis can be exported to a comma-separated values (.csv) file when the 'Export as .csv' button is clicked. The .csv file can then be opened in a text editor or spreadsheet application.

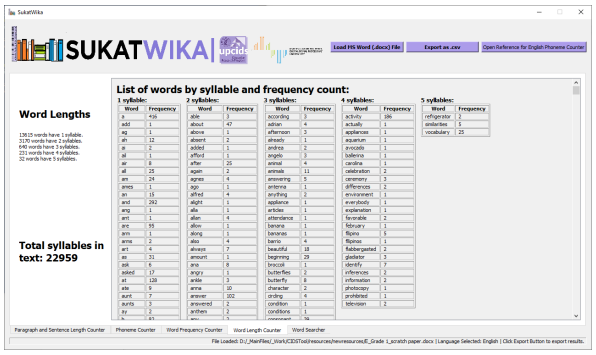


Figure 2: Tab of the word length counter showing the results of the analysis of SukatWika

An added feature that serves as an aid to users is the Phoneme Reference Guide for the English text analysis

based from the CMU Pronouncing Dictionary. This can be accessed by clicking the 'Open Reference for English Phoneme Counter' button found in the upper right portion of the user interface.

Version 1.0 of SukatWika can only support Microsoft Word (.docx) files as input and is compatible with the Windows Operating System. All dependencies are already included in the distribution of SukatWika v1.0.

3. Evaluation of Texts

Learner's materials specific for each language and grade level were gathered and collated into one file. These files were then uploaded into the program for analysis with the analysis results discussed in the succeeding section. Filipino and English versions of the tools were developed first and will serve as the basis for the Sinugbuanong Binisaya and Ilokano versions. Table 1 shows the number of learner's materials analyzed which were downloaded from the DepEd Learning Resources Portal. They are a varied collection of learner's materials made by DepEd and its partners.

	Filipino	English
Kinder	9	No texts available[1]
Grade 1	20	6
Grade 2	17	12
Grade 3	16	12

Table 1: Texts analyzed per grade level

3.1. Text Analysis Results

Syllable counter results for both English and Filipino show the progression of the length of words encountered by learners as they moved from one grade level to another. Table 2 shows the results generated by SukatWika in terms of the number of 1- to 3-syllable words in Kindergarten to Grade 3 texts.

Table 2 shows that learners in all grade levels are mostly exposed to 1-syllable words for both Filipino and English texts. The table also shows that the number and percentage of 2- and 3-syllable words increase substantially as learners go on to the next grade levels, signifying increasing complexity of texts.

This is supported by the data from the phoneme counter presented in Tables 3 and 4. Though the tables only show a specific portion of the results, the increasing frequencies of the phonemes from Kindergarten to Grade 3 can still be clearly observed, confirming the increasing level of text complexity shown by the syllable counter data.

The same trend of increasing text complexity can also be observed in the sentence and paragraph length counter results. Table 5 shows part of the sentence length counter results while Table 6 shows part of the paragraph length counter results.

[1]The mother tongue is the mode of teaching and learning in Kindergarten.

3.2. Using Text Analysis Results

3.2.1. Development of the Multi-literacy Assessments for Filipino Learners

Though an increase in text complexity is expected as learners moved from one grade level to another, SukatWika results provide the necessary details to make sound decisions on items to be included in the assessment tools. For example, the results of the phoneme counter, word frequency counter, and word length counter influenced the choice of words to be included in phonics and word reading, and spelling assessments. It enabled the assessment materials to provide a progression of word length and complexity based on the data generated by the SukatWika analysis. In identifying the list of words for the word reading test, the frequency counter results served as the basis for inclusion into the assessment tools. When listing the words for the word reading assessment, the syllable counter results and phoneme counter results validated the words that were included in the test. For example, the phoneme /ng/ was excluded from the Kindergarten assessment tools because SukatWika analyses showed that this phoneme occurred more frequently in Grade 1 materials indicating that learners had more experience with this phoneme at that level (See Table 3). It would not have been judicious to include the phoneme /ng/ in Kindergarten tools because the learners at this level can be assumed to have insufficient exposure to it in printed texts.

3.2.2. Other Uses

Sentence and paragraph length counter results may also be used as criteria for those who may want to create stories for a specific grade level. Story writers would simply need to write within the target grade level's analysis results to ensure that target readers will be able to read the text accurately.

SukatWika may also be utilized to determine an existing text's readability level. Existing stories and other learner's materials may be uploaded into the tool and the results of the analysis could then be compared with the results generated from the DepEd learner's materials to establish the material's reading level.

When planning spelling and reading lessons, teachers could also use the word search capability of SukatWika to generate words with the specific consonant blends, digraphs, or phonograms which they are studying in class. This may be especially helpful for reading remediation classes.

4. Conclusion

SukatWika enabled the development of assessment tools based on text and word properties that learners encountered in school. Since the tools are drawn from materials which learners use in school, it will yield accurate assessment results based on the exposure of learners to the printed materials.

Aside from those listed in this paper, SukatWika will have many other possible uses for teaching and assessment. It will be useful for many contexts and will hopefully open more opportunity to support reading development of all Filipino learners.

	Filipino			English		
	1-syllable words	2-syllable words	3-syllable words	1-syllable words	2-syllable words	3-syllable words
Kinder	2,999	1,582	871	No texts available[1]		
Grade 1	11,230	6,250	4,118	13,487	3,041	612
Grade 2	23,503	13,969	9,384	14,689	3,627	776
Grade 3	29,482	23,128	13,813	64,549	20,274	5,494

Table 2: Number of 1- to 3-syllable words in Kindergarten to Grade 3 Filipino and English texts generated by SukatWika

	Filipino							
	Kindergarten		Grade 1		Grade 2		Grade 3	
	Phoneme	Frequency	Phoneme	Frequency	Phoneme	Frequency	Phoneme	Frequency
Consonant phonemes	n	2,432	n	7,907	n	23,276	n	32,618
	t	1,494	ng	6,635	ng	16,045	ng	21,887
	l	1,334	s	5,402	t	13,181	t	17,790
	s	1,288	t	4,741	s	12,455	s	17,709
	k	1,208	l	4,304	m	11,092	l	15,891
Vowel phonemes	a	7,480	a	25,115	a	69,417	a	89,617
	i	1,942	i	8,225	i	21,404	i	32,624
	u	954	o	3,811	o	11,295	o	15,877
	o	868	u	3,280	u	10,342	u	12,449
	e	323	e	1,380	e	2,805	e	4,863

Table 3: Phoneme counter results for Filipino texts

Potential improvements to the software include support for additional Philippine languages, text normalization for special characters such as numbers and mathematical symbols, language identification for bilingual texts, and compatibility with other input file formats and operating systems.

5. Acknowledgements

This project is an initiative of the Education Research Program of the UP Center for Integrative and Development Studies, and is developed in partnership with the Digital Signal Processing Laboratory of the UP Electrical and Electronics Engineering Institute. We would like to thank Mr. Michael Gringo Angelo Bayona and Mr. Crisron Rudolf Lucas for their assistance in the development of the program. We would also like to thank Ms. Junette Fatima Gonzales for her input on the program and user interface and her help in translating the abstract.

6. Bibliographical References

Akademiyang Bisaya. (2011). *Cebuano Phonetics and Orthography*. Author, Cebu.

Almario (Ed.). (2014). *KWF Manwal sa Masinop na Pagsulat*. Komisyon sa Wikang Filipino, Cebu.

Carnegie Mellon University, (2014). *Carnegie Mellon Pronouncing Dictionary, version 0.7b*. Retrieved from `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`.

Cordero, G. (2019). The basic education research agenda of the department of education and system assessment in the k to 12 basic education curriculum. In D. Ocampo et al., editors, *Key Issues in Curriculum, Assessment, and ICT in Basic Education*, pages 11–30, Quezon City. University of the Philippines Center for Integrative and Development Studies.

Cummins, J. (2003). Bilingual education: Basic principles. In A. Housen J. M. Daelewe et al., editors, *Bilingualism: Language and Cognition*, pages 56–66, England. Multilingual Matters.

Dahm, R. and De Angelis, G. (2018). The role of mother tongue literacy in language learning and mathematical learning: Is there a multilingual benefit for both? *International Journal of Multilingualism*, 15:194–213.

Department of Education, (2019). *Policy Guidelines on the K to 12 Basic Education Program*. Retrieved from `https://www.deped.gov.ph/wp-content/uploads/2019/08/DO_s2019_021.pdf`.

Komisyon sa Wikang Filipino. (2012). *Tarabay iti Ortograpia ti Pagsasao nga Ilokano*. Author, Manila.

Lin, A. and Man, E. (2009). *Bilingual Education: Southeast Asian perspectives*. Hong Kong University Press, Hong Kong.

Malone, K. (1957). Syllabication. *College English*, 18(4):202–207.

Monje, J. D., Orbeta, A., Francisco-Abrigo, K., and Capones, E., (2019). *'Starting where the children are': A process evaluation of the mother tongue-based multilingual education implementation*. Retrieved from `https://pidswebs.pids.gov.ph/CDN/PUBLICATIONS/pidsdps1906.pdf`.

Nguyen, M. (2017). Bilingual education helps to improve the intelligence of children. *World Journal of English Language*, 7:11–17.

	English					
	Grade 1		Grade 2		Grade 3	
	Phoneme	Frequency	Phoneme	Frequency	Phoneme	Frequency
Consonant phonemes	t	3,756	t	4,256	t	22,881
	n	3,568	n	3,979	n	21,540
	d	2,574	d	2,937	r	16,012
	l	2,398	s	2,750	d	14,284
	s	2,310	m	1,856	l	13,380
Vowel phonemes	i (ink)	2,693	i (ink)	3,160	i (ink)	17,016
	e (modest)	1,938	a (apple)	1,828	e (modest)	11,224
	a (apple)	1,626	a (align)	1,585	a (align)	8,747
	a (align)	1,477	e (egg)	1,498	a (apple)	8,693
	e (egg)	1,456	o (button)	1,103	e (egg)	6,898

Table 4: Phoneme counter results for English texts

	Filipino			English		
	Sentences with 2 words	Sentences with 3 words	Sentences with 4 words	Sentences with 2 words	Sentences with 3 words	Sentences with 4 words
Kinder	82	89	155	No texts available[1]		
Grade 1	562	580	303	832	614	515
Grade 2	967	483	565	592	358	491
Grade 3	873	893	763	1,880	2,107	1,512

Table 5: Sentence length counter results

	Filipino			English		
	Paragraphs with 2 sentences	Paragraphs with 3 sentences	Paragraphs with 4 sentences	Paragraphs with 2 sentences	Paragraphs with 3 sentences	Paragraphs with 4 sentences
Kinder	138	22	9	No texts available[1]		
Grade 1	989	144	68	1,698	51	37
Grade 2	2,276	435	125	912	125	76
Grade 3	1,607	586	281	3,026	80	257

Table 6: Paragraph length counter results

UNESCO, (2004). *The importance of mother tongue-based schooling for educational quality*. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000146632.

UNESCO, (2018). *MTB MLE resource kit: Including the excluded: Promoting multilingual education*. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000246278.

Proceedings, pages 36–39
Paris, UNESCO Headquarters, 5-6 December, 2019.

GeTa - A Tool for the Controlled Semi-Automatic Multilevel Annotation of Classical Ethiopic

Cristina Vertan
University of Hamburg
VogtKölln Strasse 20, 22527 Hamburg
cristina.vertan@uni-hamburg.de

Abstract

Preservation of the cultural heritage by means of digital methods became extremely popular during last years. After intensive digitization campaigns the focus moves slowly from the genuine preservation (i.e digital archiving together with standard search mechanisms) to research-oriented usage of materials available electronically. This usage is intended to go far beyond simple reading of digitized materials; researchers should be able to gain new insigts in materials, discover new facts by means of tools relying on innovative algorithms.In this article we will describe the workflow necessary for the annotation of a dichronic corpus of classical Ethiopic, language of essential importance for the study of Early Christianity

Rezumat (Romanian)

Conservarea patrimoniului cultural prin intermediul metodelor digitale a devenit extrem de popular în ultimii ani. După campaniile intensive de digitizare, atentia cercetatorilor se deplasează încet de la conservarea autentică (adică arhivare digitală împreună cu mecanisme de căutare standard) la utilizarea orientată spre cercetare a materialelor disponibile pe cale electronică. Această utilizaredoreste sa patrunda mult dincolo de simpla citire a materialelor digitizate; Cercetătorii ar trebui să poată descoperi noi elemente, prin intermediul instrumentelor care se bazează pe algoritmi inovativi. În acest articol vom descrie fluxul de lucru necesar pentru adnotarea unui corpus diacronic in Ge'ez, limba etiopiana clasica , o limba de importanță esențială pentru studiul creștinismului timpuriu.

Keywords: annotation, classical Ethiopic, south-semitic language3

1 Introduction

Although of major importance for the understanding of Christian Orient, the Gəʿəz language was up to now somehow neglected by the new research directions in Digital Humanities. Substantial material in digital form exist, but there are no tools which allow a deep analysis of the language and the content.

Improving our knowledge of the Gəʿəz language is crucial in order to refine our philological and text-critical methods as well as for advancing our understanding of thought and literature expressed in Gəʿəz.

This implies a substantial enlargement of the data by:

- seizing Classical Ethiopic texts in digital form
- adding significant linguistic information
- collecting metadata
- providing tools to interpret all this information.

The project TraCES[1] (From Translation to Creation: Changes in Ethiopic Style and Lexicon from Late Antiquity to the Middle Ages) aims to fill this gap by providing a collection of reliable and extensive linguistic data based on annotated of diachronic corpus of Gəʿz. The annotation and the developed tools will enable analysis at the level of lexicography, morphology and style. The annotated texts belong to different periods and genres of Ethiopic literature (text-critical editions). The project employs a multidisciplinary approach, involving methods from linguistics, philology and digital humanities. Major results expected to bring Gəʿəz in the digital era are:

- a (deep) annotated corpus linked with
- a lexicon (first digital lexicon for Gəʿəz)
- tools for the annotation, analysis, and visualization of the corpus, and browsing the lexicon.

In this paper we will focus on the description of the annotation tool. We will explain the requirements and the challenges these requirements imply for the tool development, and we will present its components, the underlying data structure as well as the linguistic -set.

2 Challenges of Gəʿəz language for digital tools

The digital annotation and analysis of any corpus, implies several steps:

- The identification of punctuation marks
- The identification of independent tokens (Tokenisation). By token we denote the smallest

[1] Funded thought the ERC Research Grant 2014-2019 (http// https://www.traces.uni-hamburg.de/about.html)

unit to which one can assign a part-of-speech (PoS).
- The division of the text in sentences.
- The construction of a linguistic tag-set (PoS + possibly attached features and their values)
- The annotation of these features as well as attaching to each word a lemma, and a link to a language lexicon

The Gəʿəz language belongs for the moment to the group of "very low resourced languages", i.e. languages which face a significant lack of resources (corpora , lexicons, terminological data bases, Thesemantic networks) and tools.(Maegard and Krauwer 2006) defines the minimum set of such resourced and tools which are necessary to insert one language on the digital map. Usually the problematic of (very) low resourced languages is solved through adaptation of existent material for other languages within the same family. In the case of Gəʿəz this is not possible due to several issues:

- Within the semitic language family the situation is better for Arabic and Hebrew. However classical variants of these languages are as well under-ressourced. The particularities of Gəʿəz writing system (alphabet, left-right writing) make impossible any adaptation
- From the point of view of the writing system Amharic seem to be the best next candidate for an adaptation. Amharic lacks itself language resources and tools. Additionally the morphological structure differs in many points from that one of Gəʿəz

There are a number of tools which claim to be language independent. These are tool developed with a statistical paradigm: very large language corpora are used and linguistic feature are learned from those. This paradigm cannot be followed for the moment for Gəʿəz as there exist no statistically relevant Corpus for classical Ethiopic Additionally machine learning methods are quite performant when the number of features to be learned is rather small. This is not the case of Gəʿəz, for which we identified over 30 0PoS (Hummet and Druskat 2017) together with various features to be annotated.

An additional challenge is the absence of an electronic dictionary (lexicon) for Gəʿəz. Usually this is the first electronic resource to be developed for a language. Lexicons give important information about the lemma, the root as well as morphological features. The TraCES project builds the lexicon and the annotated corpus in parallel. This means that there is a bidirectional link between these 2 resources: already existing lemmas are marked in the lexicon but also new found words from the corpus are inserted (together with lemma and morphological information) into the lexicon.

A fully automatic annotation process is therefore for Gəʿəz impossible at this stage. We adopt a 2-stage workflow:

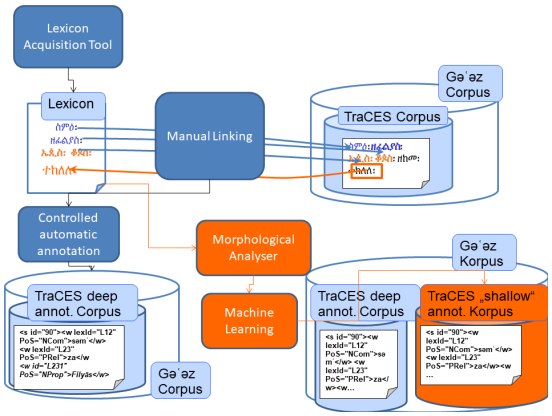


Figure 1 TraCES Modules for linguistic annotation

1. In a first stage a manual deep-annotated corpus is built. The manual Annotation is speeded-up by a controlled semi-automatic component, which will be explained in section 3
2. In a second stage the deep annotated corpus will be used as training material for a machine learning algorithm.

The complete architecture, including also the links to the lexicon component is presented in figure 1.

During last years several language-independent, respectively language customizable annotation tools were made available for researchers in humanities. Among those the most used are WebAnno (de Casthilo et. al 2014)) and CorrA (Bollmann et al. 20174) However a certain specificities of Gəʿəz made not possible the usage of these tools. In this section we will list these specificities and explain how they influenced the decisons taken for Annotation.

i) PoS Tagset

As mentioned the final goal of the TraCES project is to provide a framework which makes possible a diachronic analysis of this language. As usually variations in language occur at the micro and not the macro level, we need to perform a deep annotation which implies: a fine-grained PoS tag-set together with very precise and detailed features for each PoS. We defined a set of 30 PoS, grouped as follows:

- Nominals
  - Nouns: Common Noun, Proper Name
  - Pronouns: Independent Personal Pronoun, Pronominal Suffix, Subject Pronoun Base, Object Pronoun Base, Possesive Pronoun Base, Demonstrative Pronoun, Relative Pronoun, Interrogative Pronoun, Pronoun of Totality Base, Pronoun of Solitude Base
  - Numerals: Cardinal Numeral, Ordinal Numeral
  - Verb

- Existentials: Existential Affirmative Base, Existential Negative Base

- Particle
  - Adverbs: Interrogative Adverb, Other Adverb
  - Preposition
  - Conjunction
  - Interjection
  - FurtherParticles: Accusative Particle, Affirmative Particle, Deictic Imperative Particle, Interrogative Particle, Negative Particle, Presentational Particle Base, Quotative Particle, Vocative Particle, Other Particle
- Foreign material
- Punctuation

The inclusion of different types of particles like Prepositions and Conjunctions or relative pronouns makes imperative a splitting of Gəʿəz word units in tokens e. g.

The word unit ዘፊልያስ፡ (*zafilyās*) will be split in ዘ፡(*za*) as relative pronoun and ፊልያስ፡ (*filyās*) as proper noun.

A more challenging issue is the annotation of pronominal suffixes which can be in fact marked just in the transliteration like in the following example:

The word unit በዓሡሩ፡ transliterated as *baʿāśuru* has the following tokens: *ba* (Preposition), *ʿāśur* (common noun) and *u (pronominal suffix)*. However the pronominal suffix u is part of transliteration of the Gəʿəz letter (ሩ). Thus an annotation of such part of part of speech can be done only on transliterations.

The linguistic annotation is just part of a more complex annotation as several layers (text structure, editorial marks, named entities like persons, places, date) some of them being more appealing if they are inserted in the original script.

The annotation tool must handle in parallel the text in its original form (fidäl) and transliteration

ii) Transliteration process

Given the motivation under i) we need for all texts their transliterated version. Time constraints make impossible a manual transliteration. On the other hand a fully automatic transliteration cannot handle (without apriori knowledge) phenomena like disambiguation of 6th grade (ə) or gemmination. There are no clear linguistic rules which could cover all cases. Moreover, even some rules my imply linguistic information, which at the moment of the transliteration is not available to the system. Unsupervised machine learning approaches (without training material) will not perform satisfactory as we do not have any big corpus in both fidäl and transliteration.

Thus the annotation tool may support a kind of controlled semi-automatic transliteration 2 stages: first a rough transliteration, based on the general accepted transliteration rules is performed automatically. I-n a second stage corrections are done in a semi-automatic manner. We will explain this in section 4.

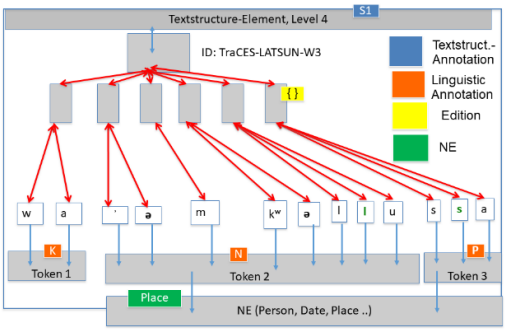
The gemmination or disambiguation of 6th grade are linguistic motivated processes. From the technical point of view the linguistic annotation is preceeded by a tokenisation process (splitting or word units in tokens). As consequence a gemmination (e.g.) may occur only after the PoS and its features are decided.

3 Underlying DataModel

The data model of the GeTa Tool follows an object-oriented approach. Each object can be located by a unique Id. There are two types of objects:

Figure 2 GeTa Data -model

Annotated Objects namely: Graphical Units, Tokens, Gəᶜz-characters and Transcription-letters.



- Annotation Objects (spans) which are attached to one or more Annotation-Objects; these are: morphological annotations, text divisions, editorial annotations.

- Links between Annotated- and Annotation-Objects are ensured through the Ids. In this way the model enables also the annotation of discontinuous elements (e.g. a Named Entity which does not contain adjacent tokens).

- A Graphical Unit (GU) represents a sequence of Gəʿz-characters ending with the Gəʿz-separator (፡). The punctuation mark (።) is considered always a GU. Tokens are the smallest annotatable units with an own meaning, for which a lemma can be assigned. Token objects are composed of several Transcription-letter objects

e.g. The GU- Object ወይቤሎ፡ contains

the 4 Gəᶜz –letter objects ; ወ, ይ, ቤ, ሎ. Each of these objects contains the corresponding Transcription-letter objects, namely:

- ወ contains the Transcription-letter objects: *w* and *a*

- ⶉ contains the Transcription-letter objects: *y* and *ə*
- ⶖ contains the Transcription-letter objects: *b* and *e*
- ⶗ contains the Transcription-letter objects: *l* and *o*

Throughout the transliteration-tokenisation phase three Token-objects are built: *wa, yəbel,* and *o*

Finally, the initial GU-Object will have attached two labels: Ⱐⶉⶖ⶗ and *wa-yəbel- o*. For synchronisation reasons we consider the word separator (፡) as property attached to the Gəᶜz-character object ⶗.

Each Token-0bject records the Ids of Transcription-letter object which he contains.

Morphological annotation objects are attached to one Token-object. They consist of a tag (the PoS e.g. Common Noun) and a list of key-value pairs where the key is the name of the morphological feature (e.g. number). In this way the tool is robust to addition of new morphological features or PoS tags.

As the correspondences between the Gəᶜz-character and the transcriptions are unique, the system stores just the labels of the Transcription-letter objects. All other object labels (Token, Gəᶜz-character and GU) are dynamically generated throughout a given correspondence table and the Ids. In this way the system uses less memory and it remains error prone during the transliteration process. In figure 3 we present the entire data model, including also the other possible annotation levels.

References

Bollmann, Marcel and Petran, Florian and Dipper,Stefanie and Krasselt, Julia 2014: *΄CorA: A web-based annotation tool for historical and other nonstandard language data΄*, in:Kalliopi Zervanou and Cristina Vertan and Antal van den Bosch and Caroline Sporleder (Eds.), Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH) Gothenburg, Sweden April 2014, 86-90.

Druskat, Stephan and Vertan, Cristina 2017, *΄ Nachnutzbarmachung von Forschungsdaten und Tools am Beispiel altäthiopischer Korpora΄,* in Gerog Vogeler (ed.) Kritik der Digitaler Vernunft Konferenzabstracts, Köln 2018, 270-273

Eckart de Castilho, Richard and Mújdricza-Maydt, Éva and Yiman, Seid Muhie and Hartmann,Silvana and Iryna and Frank, Anette and Biemann, Chris 2016, ΄A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures΄, in Erhard W. Hinrichs and Marie Hinrichs,and Thorsten Trippel (eds.), *Proceedings of the LT4DH workshop at COLING 2016*, Osaka, Japan: 76-84.

Hummel, Susanne and Wolfgang Dickhut 2016. ΄A part of speech tag set for Ancient Ethiopic΄, in Alessandro Bausi and Eugenia Sokolinski, eds, *150 Years after Dillmann΄ s Lexicon: Perspectives and Challenges of Gəʿəz Studies*, Supplement to Aethiopica, 5 (Wiesbaden: Harrassowitz Verlag, 2016), 17–29.

Krzyżanowska, Magdalena 2017. ΄A Part-of-Speech Tagset for Morphosyntactic Tagging of Amharic΄, *Aethiopica*, 20 (2017), 210–235.

Maegaard, Bente and Krauwer, Steven and Choukri, Khalid and Jørgensen, Lars, 2006, ΄The BLARK concept and BLARK for Arabic΄, in *Proceedings of the LREC Conference 2006*, http://lrec-conf.org/proceedings/lrec20

Proceedings, pages 40–42
Paris, UNESCO Headquarters, 5-6 December, 2019.

FarSpeech: Arabic Natural Language Processing for Live Arabic Speech

Naassih Gopee, Mohamed Eldesouki, Ahmed Ali, Kareem Darwish
Qatar Computing Research Institute
Hamad Bin Khalifa University, Doha, Qatar
{ngopee, mohamohamed, amali, kdarwish}@hbku.edu.qa

Abstract

This paper presents FarSpeech, QCRI's combined Arabic speech recognition, natural language processing (NLP), and dialect identification pipeline. It features modern web technologies to capture live audio, transcribes Arabic audio, NLP processes the transcripts, and identifies the dialect of the speaker. For transcription, we use QATS, which is a Kaldi-based ASR system that uses Time Delay Neural Networks (TDNN). For NLP, we use a SOTA Arabic NLP toolkit that employs various deep neural network and SVM based models. Finally, our dialect identification system uses multi-modality from both acoustic and linguistic input. FarSpeech presents different screens to display the transcripts, text segmentation, part-of-speech tags, recognized named entities, diacritized text, and the identified dialect of the speech.

Keywords: Arabic, Automatic Speech Recognition, Natural Language Processing

Resume

نقدم في هذه الورقة البحثية عرضا لدمج تقنيات المعالج الآلي للغة العربية المسمى آفراسةْ مع تقنيات آكاتش التي تحول النص المنطوق إلى نص مكتوب بشكل آني ودقيق. ولقد تطوير كل من فراسة وكاتس في معهد قطر لبحوث الحوسبة، ويعدان الأفضل من حيث الدقة من بين جميع الأدوات المتاحة. لقد أطلقنا على هذا العرض اسم آفارسبيتْش، وهذا الاسم يعد نحتا لكملتي آفراسةْ وكلمة آصوتياتْ باللغة الإنجليزية. تتميز اللغة العربية بعدة سمات تزيد من صعوبة التعامل مع المدخل الصوتي للغة، ومن أبرز هذه السمات تعدد اللهجات والتي تختلط في الكلام باللغة العربية الفصحى. في أول خطوة، يقوم النظام بتحديد لهجه المتحدث من خلال استخدام الملامح الصوتية والملامح اللغوية للكلام المنطوق، ويتم تحديدها كواحدة من خمس لهجات وهي العربية الفصحى، واللهجة الخليجية، واللهجة المصرية، واللهجة الشامية، واللهجة المغاربية. في الخطوة الثانية يقوم النظام بتحويل الكلام المنطوق إلى نص مكتوب بشكل آلي وآني وذلك باستخدام خوارزميات التعليم العميق. الخطوة الثالثة يقوم النظام باستخدام خوارزميات فراسة، والتي تقوم بتشكيل النصوص بشكل آلي وكذلك تقطيع الكلمات لمعرفة السوابق واللواحق والتعرف على أسماء الأعلام كالأماكن والمؤسسات والأشخاص. يمكنكم تجربة هذا النظام الآن من خلال الرابط التالي:
http://farspeech.qcri.org/

1. Introduction

Downstream processing of transcribed speech can enable a variety of applications such as information extraction and machine translation. In this paper, we present the FasSpeech web application, which couples QATS, our live Arabic Automatic Speech Recognition (ASR) system(0), and Farasa, our natural language processing pipeline (0), to perform a variety of NLP tasks on live speech. FarSpeech is designed to show more than the plain text speech transcripts. Farspeech has five different screens. The first four screens show the output of different Farasa NLP processors, namely; (1) Segmentation, which involves breaking words into its constituent prefix(es), stem, and suffix(es) and is essential for a variety of applications such as text retrieval and machine translation; (2) Part-of-Speech (POS) tagging, where we color code POS tags such as nouns, verbs, and adjectives using different colors; (3) Named-entity recognition (NER), which builds on segmentation and POS tagging to identify named entities that include persons, organizations, and locations; and (4) Text Diacritization, which involves the automatic recovery of short-vowels, a.k.a. diacritics, that are typically omitted in Arabic text, including in speech recognition output. ASR complicates NLP processing, because ASR output may contain recognition errors and the recognized text may contain a mix of Modern Standard Arabic (MSA) and dialects. Farasa is optimized for MSA. The final screen shows Arabic Dialect Identification (ADI) results, which attempts to label the input speech as either MSA, Egyptian, Gulf, North African, or Levantine. This is the first system that combines ASR, NLP and dialectic identification in a live streaming setup.

2. System architecture

The Farspeech system is composed of four independent components, namely: the web application, the QATS ASR server, the Farasa NLP toolkit application server, and the dialect identification system. The complete system workflow diagram is shown in Figure 1. It performs the following steps:

- Capture input from a user's microphone through the interface.
- Spawn a worker that sends raw audio to the QATS ASR server.
- Get transcribed output from QATS and send it to Farasa web server.

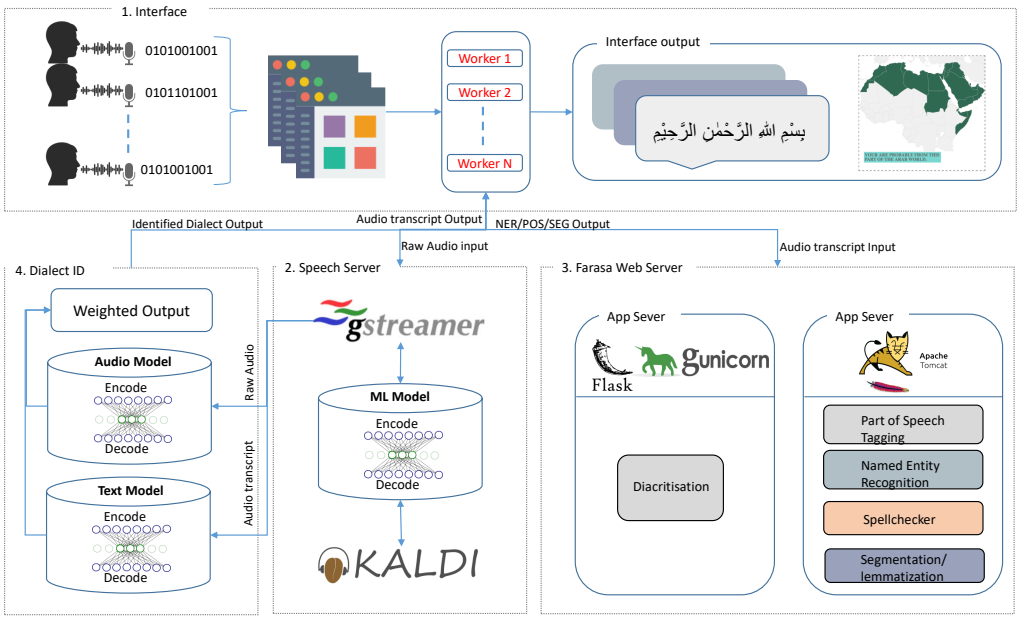


Figure 1: FarSpeech System Overview

- Retrieve Farasa results and display segmentation, POS tags, recognized named entities, and diacritized text.
- Send audio to dialect identification systems.
- Display output on Farspeech.

Since this is a live speech system, steps 1-6 are continually repeated as more audio input is received from the user. The data stream is continuously sent to the ASR server, and the incremental output is fetched to the Farasa web API for further processing allowing the different components to operate independently of each other.

2.1. The Web Application

The web application is comprised of a front-end and a back-end. The front-end presents the users with an interface to initialize their microphone and to record their audio input with the possibility of cancelling their session at any point. The front end is primarily built with the *Bootstrap*[1] framework. To handle speech we used *Wavesurfer*[2] and the *Dictate*[3] javascript library to enable microphone initialization and audio capture. *Dictate* converts the audio from WAV to raw binary audio file, which is then passed to the web application back-end that is built using the *Tornado Web Server*[4]. The Tornado application spawns workers which take care of each user's session. A worker passes the raw audio input to the ASR server and receives back the transcript as an output. Once the transcribed output is received, the data is sent to the Farasa web server through its web API, which returns processed text, including segmentation, POS tags, named entities, and diacritization. The raw audio together with its text transcripts is also sent in parallel to the dialect identification system which in turns return the mostly likely region from which the dialect originated. These steps are repeated for each partial audio input that the application receives from the user microphone input.

2.2. Speech transcription

We use the QATS Speech-to-text transcription system that we built as part of QCRI's submission(0) to the 2016 Arabic Multi-Dialect Broadcast Media Recognition (MGB) Challenge(0). The key features of the transcription system are as follows:

Acoustic Models: We experimented with various acoustic models; Time Delayed Neural Networks (TDNNs) (0), Long Short-Term Memory Recurrent Neural Networks (LSTM) and Bi-directional LSTM (BiLSTM) (0). Though the performance of the BiLSTM acoustic model in terms of *Word Error Rate* is better than the TDNN, TDNN has a much better *real-time* factor while decoding. Therefore, we opted to use the TDNN acoustic model. More model details are available in Khurana et al. (0).

Language Model: We built a Kneser Ney smoothed trigram language model. The vocabulary size is restricted to the 100k most frequent words to improve the decoding speed and in-turn the *real-time factor* of the system. The choice of using a trigram model instead of a recurrent neural network model, as in the QATS offline system, was essential for keeping the decoding speed at a reasonable fast.

2.3. Natural language processing

For Arabic text language processing, we used our in-house Arabic NLP toolkit called Farasa[5] (meaning chivalry in Arabic). The pipeline includes segmentation, POS tagging,

[1] https://getbootstrap.com/
[2] https://wavesurfer-js.org/
[3] http://kaljurand.github.io/dictate.js/
[4] https://www.tornadoweb.org/
[5] http://farasa.qcri.org

NER, a diacritization, spell checking, lemmatization, and dependency and syntactic parsing. Though Farasa is tuned for MSA, particularly for the news domain, it can handle other genres along with classical and dialectal Arabic, but at reduced accuracy. This is possible because of the large overlap between MSA and other varieties of Arabic. Farasa fills an important gap in the span of available tools. It is the only comprehensive suite of Arabic tools that is both open source and whose internal sub-components are competitive with the state of the art. In FarSpeech, we employ the segmenter, POS tagger, NER, and diacritizer.

3. Conclusion

This paper presents FarSpeech, the QCRI system for live speech, NLP processing, and dialect identification. Currently, the system works very well for Arabic including frequent dialectal words. For future work, we aim to improve the system in several ways including having a tighter integration between ASR and NLP, and to extend its use to other applications such language learning.

A. Ali, Y. Zhang, and S. Vogel, "Qcri advanced transcription system (qats)," in *Spoken Language Technology Workshop (SLT)*, 2014.

A. Abdelali, K. Darwish, N. Durrani, and H. Mubarak, "Farasa: A fast and furious segmenter for arabic," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 2016, pp. 11–16.

S. Khurana and A. Ali, "QCRI advanced transcription system (QATS) for the arabic multi-dialect broadcast media recognition: MGB-2 challenge," in *Spoken Language Technology Workshop (SLT) 2016 IEEE*, 2016.

A. Ali, P. Bell, J. Glass, Y. Messaoui, H. Mubarak, S. Renals, and Y. Zhang, "The mgb-2 challenge: Arabic multi-dialect broadcast media recognition," in *SLT*, 2016.

V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts." in *INTERSPEECH*, 2015, pp. 3214–3218.

H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling." in *INTERSPEECH*, 2014.

Proceedings, pages 43–46
Paris, UNESCO Headquarters, 5-6 December, 2019.

Handling Prosody and Tone Languages

Aijun LI, Wei WANG
Institute of Linguistics, Chinese Academy of Social Sciences, Beijing, China
No.5 Jianguomennei Dajie, Beijing, 100732, P.R.C.
liaj@cass.org.cn, waywon@qq.com

Abstract

Technology has played a vital role in advancing our understanding of prosody in human languages. In many languages, tones distinguish lexical meanings and variations of other prosodic features carry semantic or pragmatic information. Although deep learning and end-to-end technologies have been increasingly used in speech applications, challenges in technology for handling prosody and tone languages are still significant, especially in less familiar languages and dialects. In this squib, we will present as a case in point the study of tone, intonation and other prosodic features in Chinese and its relevance to speech technology.

Keywords: tone, intonation, prosody, prosodic modelling

Résumé

技术在我们对了解人类不同语言中起了非常关键的作用，语言学家采用科学仪器研究声音的声学、发音和感知特性，包括韵律和声调特性。声调可以区分词义，韵律特征的变化可以传递言语的语义和语用信息。对言语技术来说，尽管已经很容易采用深度学习和端到端技术来构建语音应用系统，面对智能语言系统来说，特别是对一些低资源的语言或者方言的研究，仍存在很多挑战，需要加强对韵律和声调语言的研究。本文将重点以声调语言汉语为例，围绕言语技术面临的这些挑战，介绍声调、语调和韵律的研究成果。

关键词: 声调，语调，韵律，韵律建模

1. Introduction

Tones can be defined as pitch variations that change either the lexical or grammatical meaning of a word. A language in which the meaning of a word depends on its tone is known as a tone language. The aspects of speech that extend beyond individual vowels and consonants are known as suprasegmentals or prosody. Narrowly speaking, prosody is sometimes used as a synonym only for intonation, which refers to the use of suprasegmental features to convey post-lexical or sentence-level pragmatic meanings in a linguistically structured way (Ladd, 1996).

Tone languages can have either register tones or contour tones. The vast majority of languages spoken in Africa are register tone languages, such as Igbo, Shona, Yoruba, and Zulu, to name just a few. In a contour tone language, pitch movement, instead of pitch level, serves to distinguish word meaning. Many of the languages spoken in Southeast Asia – including Mandarin Chinese, Chinese dialects, Thai, and Vietnamese – are contour tone languages.

Speech prosody is a systemic structure of various linguistic components in an utterance or multiple connected utterances, which conveys linguistic, paralinguistic and even nonlinguistic information via suprasegmental features (Fujusaki, 1997). As a result, prosody is not only related to tone, intonation, stress and rhythm, but also to discourse-information, as well as interlocutor's emotions and attitudes.

Deep learning and end-to-end technology are widely used in speech application systems. However, challenges are plentiful for technology in handling prosody and tone languages in real-life intellengent systems, to wit, how to describe complex tonal systems phonetically and phonologically; how to model tone and intonation from a typological point of view; how to model contextual prosody in interaction; and how to apply technology in prosodic annotations for speech corpora with different accents automatically or semi-manually.

To this end, in this squib I will focus on presenting as a case in point the study of tone, intonation and other prosodic features in Chinese, including phonetic and phonological descriptions of tones, tone sandhi and tonal coarticulation, typological study on intonation, intonation modeling and prosodic annotations.

2. Phonetic and Phonological Despcriptions of Tones

In the first comprehensive explication of the sound system in Chinese found in a Chinese rhyme dictionary called "Qieyun", created in 601 during the Sui Dynasty (581–618), a total of 12,000 character entries were arranged corresponding to the four tonal categories, referred to as "Ping", "Shang", "Qu", and "Ru" in the philological tradition. A new phase was initiated in the experimental study of tone when the kymograph was used in the measurement of tones in Chinese dialects by Liu Fu, Wang Li and Yuen-Ren Chao. Chao was also the inventor of "tone letters" used in International Phonetic Alphabets (IPA) for the transcription of tones, in which the total pitch range is divided into four equal parts, marked by five pitch levels, numbered 1, 2, 3, 4, 5, with 1 being the lowest pitch and 5 the highest (Chao, 1930).

Presently, there are ten major dialect groups of Chinese spoken in China (Maps of the languages in China, 2012), which exhibit great variations in phonology, syntax and other aspects of grammar.

Southern dialects generally have more complex tonal systems than northern dialects in terms of the number of tones and contour shapes. While most southern dialects have between 6 and 13 tones, northern dialects have about 2 to 5 tones. Mandarin dialects are not known for having too many tones, except those spoken in the lower reaches

of Yangtze River. Most of them have 5 to 7 tones, due to partly influences from the neighboring Wu dialects. On the other end of the spectrum, the Lanyin variety of Mandarin, which is spoken in the western part of China, has 2 to 4 tones, fewer than most other Mandarin dialects due to its direct contact with non-tonal minority languages over a long period of time. According to Ran (2018), the dialect with the most tones is a Gan dialect called the Jinxian dialect, which has 13 tones; and the dialect with the fewest tones is the Honggu dialect, with only 2 tones.

One area focused on the analysis of acoustic features such as f0, duration and amplitude of tones in isolation and in a sequence (e.g. Lin and Yan, 1992; Wu, 1982). An interesting development in recent years has been the study of phonation types such as modal voicing, breathy voice or creaky voice. Evidence was presented to show that lexical meanings can be distinguished on the basis of both pitch and phonation types, not just pitch alone as previously assumed for tone languages like Chinese (Kong, 2001; Zhu, 2012).

Phonological systems to represent tones have been proposed for Mandarin and other Chinese dialects. Two earliest studies (Wang, 1967; Wu, 1979) both distinguished register and contour features of tones. The earliest study concerning the representation of tones is Wang (1967). He distinguished three level features and a redundant contour feature. For example, a tone can be in the high or low register of pitch and it can take the shape of a level or contour pitch movement such as rising or falling. The dichotomy of register and contour is widely adopted in more formal treatment of representation of tones (Bao, 1990; Chen, 2000; Duanmu, 1990; Li, 2003; Yip, 1989, 2000). In a comprehensive study of tonal systems, Liu (2004) surveyed 1,186 Chinese dialects from published documentations. Tones in her study are transcribed with the tone letters system of Chao. She proposed register and contour feature to classify tone systems of Chinese dialects and accounted for contour variations using downtrend of pitch.

Zhu (2012) departed from previous systems of tonal features and proposed a new notational system, based on his own field work. According to him, in Chinese dialects some tones are not only cued by pitch movement, but also by phonation types. Therefore, phonation types and pitch movement are both relevant in defining three pitch registers and six pitch levels. He proposed a framework of tonal typology in which each tone can be defined by four features which are register, duration, level, and contour.

3. Tones in Context: Tone Sandhi and Tonal Coarticulation

Changes can happen as a result of two tones in juxtaposition. Tone sandhi is generally known as the process of tone change due to the influence of the other. The best-known tone sandhi in Mandarin is the third tone sandhi: in the combination of two syllables in Tone 3, the first syllable changes to Tone 2. In other words, Tone 3 + Tone 3 become Tone 2 + Tone 3. It is a macro change of tonal identity and is easily noticeable by fluent speakers. Tonal coarticulation, on the other hand, refers to much more subtle changes, caused by interactions of different articulatory gestures. It does not involve changes of tonal categories.

Major issues in the study of tone sandhi include the relationship between tone in isolation and in tone sandhi, the domain in which tone sandhi happens and the types of tone sandhi. Tone sandhi patterns can be classified in terms of direction, manifestation and domain. For example, the Shanghai dialect and other Northern Wu dialects exhibit the left-dominant tone sandhi pattern in which the left-most syllable in the tone sandhi domain – lexical compounds in this case – keeps its underlying tone and spreads its tone to the other syllables in the domain. Phrasal structures follow different patterns, though. The right-dominant pattern is illustrated in the third tone sandhi in Mandarin above: The Tone 3 in the right-most syllable triggers the tone sandhi in the previous syllable. A third pattern is called tonal substitution. An often-cited example is the southern Min dialect spoken in Xiamen (also known as Amoy). The right-most syllable keeps its tone and the other syllables in the same domain are replaced with other tones in a circular way, forming what is known as the Min circle (Chen, 1987). Chen (2000) provides in-depth analyses of tone sandhi patterns across Chinese dialects. It has been so far the most comprehensive study of tone sandhi in Chinese dialects.

Determination of tone sandhi domain has been challenging as it involves syntax, rhythm and tempo. Take the third tone sandhi in Mandarin as an example. The tone sandhi patterns can be quite complex when three or more third tones are present in the same domain and across two domains. As mentioned above in passing, compounds and phrasal structures in Wu dialects follow different set of tone sandhi rules.

Formal analyses of tone sandhi are often given in tonal features such as H (high) and L (low). Different theories have been proposed for tone sandhi patterns in different dialects, but there are still remaining issues that have yet to be fully accounted for.

Tonal coarticulation happens between adjacent tones, but it does not cause changes of tonal categories. Effects of tone sandhi in Mandarin have received intensive attention since the 1990s (Peng, 1997; Shen, 1990; Shih, 1986; Xu 1994, 1997, 1999). Both carryover and anticipatory effects have been identified in the production of connected tones. For example, in Mandarin when the Tone 3, a low tone, is preceded by the other three tones and itself, the pitch trajectory of the low tone is greatly affected by the preceding tones. Specifically, its high point is the highest at the syllable boundary when preceded by a high level tone. The carryover effect in Mandarin is primarily assimilatory. The anticipatory effect, much smaller in scale than the carryover effect, is seen in the raising effect of the low tone on the preceding tone. For example, the high f0 (fundamental frequency) of the falling tone in Tone 4 is realized higher after Tone 3 than the other three tones, ergo the effect is mainly dissimilatory.

In addition to four tones in Mandarin, there is a special category called the neutral tone. A syllable is said to be in neutral tone when it does not carry one of the four lexical tones in Mandarin. Neutral tone has received ample attention in research, but its status in the Mandarin phonological system has been controversial as to whether the neutral tone is a tonal category, or it is related to stress. According to Chao (1979), neutral tone is related to weak stress. When a syllable is in weak stress, its tonal range is almost reduced to zero and its duration significantly

shortened. Lin (1957, 1962) used the term "weak stress" to capture the metrical property of syllables in neutral tone. Lu and Wang (2005) distinguished "neutral tone" from "weak stress", reserving the former as neutralized tone in the tonal system and the latter as unstressed syllable in the metrical system. Their view resonates with Chao's. Li (2017) approached the neutral tone by considering the effects of prosodic boundary and information structure in the acoustic analysis, and examined its pitch and durational properties in different prosodic contexts.

4. Typological Study on Intonation

After the emergence of the prosodic or intonational annotation system such as INTSINT (Hirst, 1998) and ToBI (Silverman et al. 1992), a number of studies have been conducted within these frameworks, in which the view of local typology is widely assumed. Intonational typology does not merely compare intonation structures of different languages, but is also relevant to theoretical issues, among which question intonation and the interplay between intonation and lexical prosody such as tone are most discussed. The difficult development of intonational typology is due to the complexity of intonation as well as the lack of a widely accepted annotation system. Moreover, the researchers still do not reach a consensus as to the parameters of cross-linguistic comparisons of intonation. Jun (2005, 2014) present analyses of intonation in 27 languages, including Mandarin, and describe how intonation is constructed out of tones or accents that are tied to stressed syllables and syllables at the end of a prosodic domain, following the autosegmental-metrical (AM) approach to intonation (Ladd, 1996, 2008).

The interplay of tone and intonation in Chinese has been most intriguing in intonational phonology. Since the seminal work of Yuen-Ren Chao (Chao, 1932, 1933), a rich body of research, both descriptive and experimental, has been produced to advance our understanding of the linguistic functions and physical properties of tone and intonation in Mandarin Chinese, especially regarding the interaction of tone and intonation. Chao came up with two metaphors that have been widely known in characterizing how tone and intonation interact: the "rubber band effect" and the "small wave and big wave" theory. According to the latter theory, tone and intonation are related in the form of superimposition – either successive or simultaneous – just like small waves sitting on top of big waves.

The study of intonation in Chinese dialects is still in its infancy, but has shown great potential in making meantinful contributiongs to the typological study of intonation. More comprehensive studies other than Mandarin can only be found in Cantonese (Fox et al., 2008). The study of Cantonese intonation started early in 1970s, when Vance investigated the tone and intonation in Cantonese (Vance, 1973). Other studies in the early days include those of the Chengdu dialect (Chang, 1958), Toishan (Lee, 1986), Cantonese (Johnson, 1986) and Changsha (Shen, 1991) dialects.

The study of Cantonese treated different aspects of intonation: the focus structure (Man, 2002), the effect of question intonation on lexical tone (Lee, 2004; Ge, 2018), the interaction between sentence final particles and intonation (Wu, 2008), and the modeling of Cantonese intonation using the command-response model (Gu, 2004). The most notable property of Cantonese intonation is the rising question intonation, as it has been demonstrated in several studies on acoustics (Lee, 2004; Ge, 2018) and perception (Mai, 2000). Another noteworthy phenomenon is its rich inventory of sentence final particles (SFPs). As is shown by Wu (2008), the pitch manifestations of SFPs combine the effects of intonation and lexical tones.

There are a few of studies of intonation in other Chinese dialects lately, such as the Tianjin dialect (Zhang, 2018), and the Kaifeng dialect (Wang, 2018). Several other studies also looked at the focus structure in other dialects, such as Teochew (Hsu et al., 2018) and the Shanghai dialect (Ling and Liang, 2017). The focus structure of some Shandong dialects has also been investigated by Jia and colleagues (Jia, 2011; Duanmu et al.2013; Duanmu and Jia, 2015). As in Mandarin, PFC is found prevalently in many Chinese dialects, except Cantonese. Cantonese distinguishes nine lexical tones (including three checked tones), and it does not show the PFC effect. Another dialect that has been shown to lack this phenomenon is Taiwanese (Chen et al., 2009). More work is in need in this area.

5. Tone and Intonation: Modeling and Beyond

Following the footsteps of Chao, many were engaged in developing models of intonation in Mandarin to account for the interaction of tone and intonation. Wu (2004), who inherited and expanded Chao's theory, proposed the "transposition model" of intonation, which accounts for obligatory and optional tone sandhi patterns in Chinese. Shen (1992, 1994) characterized intonation in terms of the upper line and lower line of f0 that define a pitch register and argued that the two lines can be manipulated independently of each other in different intonation patterns. Xu (2004) proposed the Parallel Encoding and Target Approximation (PENTA) model of speech prosody, which is a framework for conceptually and computationally linking communicative meanings to fine-grained prosodic details, based on an articulatory-functional view of speech.

In a monograph on the experimental study of tone and intonation in Chinese, Lin (2012) took Chao's insights as a point of departure and explicitly adopted the autosegmental-metrical (AM) model of intonation. In his model, focal prominence and boundary tone are the two key elements in describing intonation in Chinese. For example, the difference between declarative intonation in statements and interrogative intonation in questions without sentence-final question particles resides in the boundary tone, which is realized acoustically as pitch register and slope of the contour. Shi (2013) looked at intonation from a broader perspective and proposed a systematic method to define an "intonation pattern" with three parameters – f0 contours, pause-lengthening ratio and sound intensity. In his study of declarative and interrogative intonations in Putonghua Cantonese and Korean, Shi was trying to figure out cross-linguistic patterns in intonation in terms of the quantifiable measurements of f0, duration and intensity.

Focus was probably given the most attention in the study of intonation. Xu (1999) looked into the effects of tone and focus on the alignment of F0 contours and found that focus exerts influence on pitch range in different ways: the pitch range of the syllables before the focal position remains

unmodulated, and that of the syllables is dramatically expanded in the focal position and compressed after focus, a phenomenon he termed "post-focus compression" (PFC). Other studies analyzed phonetic realizations of different types of focus and situations in which there are one, two or multiple foci in the utterance (Jia and Li, 2012). Wang and Xu (2011) reported an experimental investigation of the prosodic encoding of topic and focus in Mandarin by examining disyllabic subject nouns elicited in four discourse contexts. Their major findings were that focus causes post-focus f0 lowering while topic allows a gradual f0 drop afterwards; the effects of downstep, sentence length on initial f0 are independent of topic and focus, and the effects of topic, focus, downstep and sentence length are largely cumulative.

In addition to focus, prosodic structure has been another closely-examined area in the study of intonation. Tseng (1999) proposed a HPG model and modelling the tone and intonation under the frame of discourse prosody.

The link between intonation and emotion was explored as early as Bolinger (1989). Experimental studies have flourished on the influence of emotions on intonation patters in recent years (e.g. Bänziger and Scherer, 2005). Li (2015) undertook an extensive study on the role of intonation in conveying emotion in a tone language like Chinese, with focus on f0 levels and pitch contours in what she termed "successive addition boundary tone". She proposed that the boundary tone is composed of two components – the base tone of the syllable and an addition contour.

Intention understanding and generating in human-machine interactions calls for greater integration of discourse-level prosodic information in spoken dialogue systems. In a series of studies on the interface of prosody and discourse, Li A., Jia and their collaborators conducted detailed analyses on prosodic features in connection with discourse structure, information structure and dialogue acts (Jia, 2018; Li, 2018; Li et al., 2019).

Prosodic information is widely used for the detection of disfluencies and utterance boundaries, the segmentation of dialogue acts, the detection of sentence mood and modality, accent and so on. Prosodic annotation system, which provides a tool to highlight significant prosodic events and which is essential to statistical prosody modeling in these tasks. The prosody annotation systems based on ToBI framework and its modifications, such as C-ToBI (Li, 2002). and KToB, are most popular. Since manual prosodic annotation, is generally time-consuming and expensive to administer. It is important to develop automatic annotation of prosodic information. Many algorithms or models were proposed: CRF and HMM for annotation of Japanese accent types and phrase boundaries (Koriyamay et al., 2014), unsupervised joint prosody labeling and modeling by Chiang et al. (2009) for read speech and spontaneous Mandarin speech by Lin et al. (2016), and transfer learning and RNN-based model for L2 prosodic annotation and evaluation (Lee, 2019; Chen, 2019)

6. Concluding Remarks

Technology for handling Prosody and tone has now become more interdisciplinary and better integrated with other disciplines than ever before. Looking forward, we expect that availability of advanced research instruments is adopted to explore speech production and perception mechanisms at the neurological level, to direct more attention and resources to cross-linguistic and cross-dialectal typological and applicational research in order to better serve diverse linguistic and dialectal groups, and to conduct studies on contextual tonal and prosodic variations to meet the demands of speech and language technology. Finally, technology in speech and language technology will continue to serve as an invaluable tool in our joint efforts to document and preserve endangered languages and dialects in order to strengthen linguistic diversity.

7. Acknowledgements

This research is supported by the Key NSSFC Granting (No.15ZDB103), the National Key R&D Program of China (No. 2017YFE0111900).

8. Selected References

Bao, Z. (1990). *On the nature of tone*. Doctoral dissertation, MIT.
Chao, Y. R. (1933). Tone and Intonation in Chinese. In *Linguistic Essays by Yuenren Chao*, Beijing: The Commercial Press.
Chao, Y. R. (1934). The Non-uniqueness of Phonemic Solutions of Phonetic System, in *Linguistics Essay by Yuenren Chao*, Beijing: The Commercial Press.
Chen, M. Y. (2000). *Tone Sandhi: Patterns across Chinese Dialects*. Cambridge: Cambridge University Press.
Chinese Academy of Social Sciences. (2012). *Atlas of Languages in China.* Beijing: The Commercial Press.
Duanmu, S. (1990). *A formal study of syllable, tone, stress and domain in Chinese languages*. Doctoral dissertation, MIT.
Hirst, D., ed. (1998). *Intonation Systems: A Survey of Twenty Languages*. Cambridge: Cambridge University Press.
Jun, S. (2005). *Prosodic Typology: The Phonology of Intonation and Phrasing*. Oxford, New York: Oxford University Press.
Jun, S. (2014). *Prosodic Typology II*. Oxford, New York: Oxford University Press.Li, A. (2002). Chinese prosody and prosodic labeling of spontaneous speech. In *Proceedings of Speech Prosody*.
Li, A. (2015). *Encoding and Decoding of Emotional Speech: A Cross-Cultural and Multimodal Study between Chinese and Japanese (Prosody, Phonology and Phonetics)* 1st ed. Springer.
Li, Z. (2003). *The phonetics and phonology of tone mapping in a constraint-based approach*. Doctoral dissertation, MIT.
Lin, M. (2012). *The experimental study of Chinese intonation*. Beijing: China Social Sciences Press.
Peng, S. H. (1997). Production and perception of Taiwanese tones in different tonal and prosodic contexts. *Journal of Phonetics*, 25(3): 371-400.
Potisuk, S., Gandour, J., and Harper, M. P. (1997). Contextual variations in trisyllabic sequences of Thai tones. *Phonetica*, 54(1): 22-42.
Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., ... and Hirschberg, J. (1992). ToBI: A standard for labelling English prosody. In *Proceedings of ICSLP*.
Shen, X. S. (1990). Tonal coarticulation in Mandarin. *Journal of Phonetics*, 18(2): 281-295.
Shih, C. L. (1986). *The Prosodic Domain of Tone Sandhi in Chinese*. Doctoral dissertation, University of California, San Diego.
Wu, Z. (2004). *Linguistics Essay*. Beijing: The Commercial Press.
Xu, Y. (1994). Production and perception of coarticulated tones. *Journal of the Acoustical Society of America*, 95(4): 2240-2253.
Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics*, 25(1): 61-83.
Xu, Y. (1999). Effects of tone and focus on the formation and alignment of F0 contours. Journal of Phonetics, 27(1): 55-105.
Xu, Y. (2004). Transmitting tone and intonation simultaneously - the parallel encoding and target approximation (PENTA) model. *Proceedings of International Symposium on Tonal Aspects of Languages*.
Yip, M. (1989). Contour tones. *Phonology*, 6(1):149-174.
Yip, M. (2002). *Tone*. Cambridge: Cambridge University Press.

Proceedings, pages 47–50
Paris, UNESCO Headquarters, 5-6 December, 2019.

Talking with Robots: Opportunities and Challenges

Roger K. Moore
Speech and Hearing Research Group (SPandH)
University of Sheffield
Regent Court, 211 Portobello
Sheffield, S1 4DP, UK
r.k.moore@sheffield.ac.uk

Abstract

Notwithstanding the tremendous progress that is taking place in spoken language technology, effective speech-based human-robot interaction still raises a number of important challenges. Not only do the fields of robotics and spoken language technology present their own special problems, but their combination raises an additional set of issues. In particular, there is a large gap between the formulaic speech that typifies contemporary spoken dialogue systems and the flexible nature of human-human conversation. It is pointed out that grounded and situated speech-based human-robot interaction may lead to deeper insights into the pragmatics of language usage, thereby overcoming the current 'habitability gap'.

Keywords: spoken language technology, human-robot interaction

Résumé

Malgré les énormes progrès réalisés dans la technologie de la langue parlée, une interaction homme-robot efficace basée sur la parole soulève encore un certain nombre de défis importants. Non seulement les domaines de la robotique et de la technologie de la langue parlée posent des problèmes particuliers, mais leur combinaison soulève un ensemble de problèmes supplémentaires. En particulier, il existe un large fossé entre le discours stéréotypé qui caractérise les systèmes de dialogue parlés contemporains et la nature flexible de la conversation homme-humain. Il est souligné que l'interaction homme-robot fondée et basée sur la parole peut mener à une compréhension plus approfondie de la pragmatique de l'utilisation du langage, surmontant ainsi le 'fossé d'habitabilité' actuel.

1. Introduction

Recent years have seen tremendous progress in the deployment of practical spoken language systems - see Figure 1. Commencing in the 1980s with the appearance of specialised isolated-word recognition (IWR) systems for military command-and-control equipment, spoken language technology has evolved from large-vocabulary continuous speech recognition (LVCSR) for dictating documents (such as Dragon's *Naturally Speaking* and IBM's *Via Voice*) released in the late 1990s, through telephone-based interactive voice response (IVR) systems to the launch of *Siri* (Apple's voice-enabled personal assistant for the iPhone) in 2011. *Siri* was quickly followed by Google *Now* and Microsoft's *Cortana*. The following years heralded a new era of smart speaker based voice assistants, starting with Amazon's 2015 release of *Alexa* followed later by *Google Home*, Apple's *HomePod* and Sonos *One*.

These contemporary systems not only represent the successful culmination of over 50 years of laboratory-based speech technology research (Pieraccini, 2012), but also signify that speech technology had finally become "mainstream" (Huang, 2002) (at least, in the English-speaking world). Indeed, the market penetration of these smartphone and smart speaker based voice assistants is astounding. For example, Siri has had over 40 million monthly active users in the U.S. since July 2017, Google Assistant is available on over 225 home-control brands and more than 1,500 devices, and tens of millions of Alexa-enabled devices were sold worldwide over the 2017 Christmas holiday season (Boyd, 2018). Also, a study by Juniper Research (Smith, 2017) estimated that the number of voice assistant devices across all platforms (smartphones, tablets, PCs, speakers, connected TVs, cars and wearables) would reach 870 million in the U.S. by 2022.

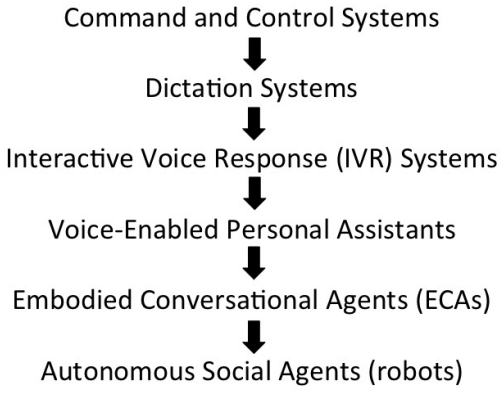


Figure 1: The evolution of spoken language processing applications from specialised military 'command-and-control' systems of the 1980/90s to contemporary 'voice-enabled personal assistants' (such as *Siri* and *Alexa*) and future 'autonomous social agents', i.e. robots.

Research is now focused on verbal interaction with embodied conversational agents (such as on-screen avatars) and autonomous social agents (such as robots), based on the assumption that spoken language will provide a 'natural' interface between human beings and future (so-called) intelli-

gent systems, and first-generation devices (such as *FurHat*[1] and *Olly*[2]) have already begun to enter the commercial marketplace.

However, notable casualties (such as *Jibo*[3] which famously announced its own demise in June 2019) confirm that there are significant challenges as well as opportunities in creating spoken language based interaction between people and robots (Moore, 2015). Some of these are discussed below.

2. Why Robots?

Before discussing the challenges of talking with robots, it is useful to recall why robots are of interest in the first place. First and foremost, developments in robotics are driven by the many benefits provided by *automation*. Since the beginning of time, humans have been inventing technologies to ease their daily toil, and the industrial revolution heralded an era of increasing automation using ever more sophisticated machines. The benefits of doing so include making/saving money, saving time and effort and improving the quality of life. Robotics - driven by the recent surge in artificial intelligence (AI) - represents the latest attempts at automation, particularly for doing things that are difficult, dirty, dangerous or dull.

3. What is a Robot?

A robot is harder to define that one might think. As Joseph Engelberger (1925-2015), developer of the first industrial robot in the United States in the 1950s, famously said: "*I can't define a robot, but I know one when I see one*"!

In fact there are a number of definitions of a robot, and the following is typical ...

> "*A robot is an actuated mechanism programmable in two or more axes with a degree of autonomy, moving within its environment, to perform intended tasks.*"[4]

They key idea is that a robot is a *physical machine* (i.e. capable of movement within in environment, whether it is real or simulated), *autonomous* (i.e. capable of acting without constant human intervention) and *programmable* (i.e. it is more than just an automaton). This means that Siri and Alexa are *not* robots (since they are incapable of moving or acting on the world), nor are tele-operated devices such as remote-controlled drones (since they are not autonomous), and nor is Terminator (since it is purely fictional!). Typical robots are thus those that one would find on an industrial production line, floor-cleaning robots (such as Roomba[5]), and humanoid robots (such as Pepper[6]).

[1] https://www.furhatrobotics.com

[2] https://www.heyolly.com

[3] https://www.jibo.com

[4] http://www.leorobotics.nl/definition-robots-and-robotics

[5] https://www.irobot.co.uk/roomba

[6] https://www.softbankrobotics.com/emea/en/pepper

4. Why Talk with a Robot?

As with all technology, there are huge benefits to be gained when humans are 'in the loop'. For example, a modern automobile already exhibits several levels of automation (e.g. power-assisted steering and cruise control) combined with human involvement in low-level activities such as acceleration and braking. As technology moves towards more autonomous vehicles and the degree of automation increases, human involvement will shift to higher levels (such as defining the destination and required time of arrival) with low-level interventions only occurring in exceptional circumstances (e.g. in an emergency). Such high-level interactions would seem to be very appropriate for a communication channel such as speech.

The field of 'human-robot interaction' (HRI) is concerned with these issues and, in particular, how to maximise the effectiveness of such interaction in a multi-modal context, e.g. vision, sound, haptics, and of special interest here, speech and language. So, how might spoken language play a role in human-robot interaction? This can be answered by considering three domains in which such interaction might take place: the *physical* world of stuff and things, the *social* world of people, agents and relations, and the *abstract* world of ideas, information, data and thought.

4.1. Speech-based HRI in the Physical World

Human-robot interaction in the physical world is often concerned with the provision of mechanical support for the human being, e.g. allowing a person to lift a heavy object or pilot a vehicle. Much of the low-level interaction could be achieved by the manual operation of physical controls and observing visual displays, but the introduction of a speech channel would facilitate additional control even if the users hands are occupied, and/or the ability to receive information even if the eyes are engaged in a more critical task (such as watching for hazards). Such activities are known as *eyes-busy, hands-busy* scenarios, and they are prime candidates for speech-based HRI.

In general, physical HRI is targeted at *collaborative working* where tasks are distributed between human and robot teams. In such situations, speech can offer a powerful means for coordinating actions ("*Pass me the wrench*") and for managing joint attention ("*Mind that hole!*").

4.2. Speech-based HRI in the Social World

Human-robot interaction in the social world is concerned with the provision of emotional and/or motivational support for the human being, e.g. through *companionship* and the exhibition of empathy or even dominance (as would be required from a personal trainer). Such behaviours would serve to underpin the relations between the different actors/agents and their individual and/or collective roles and responsibilities.

In general, social HRI would exploit both verbal and non-verbal channels of communication, and would naturally draw on the expressive *paralinguistic* properties of spoken language.

4.3. Speech-based HRI in the Abstract World

Human-robot interaction in the abstract world is concerned with the provision of mental support for the human being, e.g. by giving access to the vast amounts of information/data available on the internet. Spoken language not only offers a more intuitive (some say 'natural') method of human-robot communication, but it also supports a very high information-rate exchange compared to that available through the physical or social channels.

5. Challenges for Speech-based HRI

5.1. Issues Arising from Robotics

There are many challenges facing the opportunities identified above. Not only are there a number of difficulties to be overcome in the core area of speech-based human-robot interaction, but problems are also inherited from the field of robotics in general. For example, all robots are complex mechanical, electrical, electronic and computer-based physical machines operating in the real world, which means that they can be very fragile. A network outage, a broken spring, or a computer bug can easily bring operations to a halt (or worse), and the likelihood of some component failing can be quite high. Also, robots tend to be quite expensive pieces of equipment, meaning that personal ownership may be challenging for particular user groups.

5.2. Issues Arising from Spoken Language Technology

Likewise, all the problems facing mainstream spoken language technology also apply to speech-based human-robot interaction. For example, strong accents, minority languages, and noisy environments can all lead to poor performance of the speech technology components which, in turn, will have a negative impact on the effectiveness of speech-based HRI.

5.3. Issues Arising from Speech-based HRI

In addition, there are many issues that arise from speech-based human-robot interaction itself. For example, robots are quite noisy, hence listening and moving are often incompatible activities[7]! Also, everyday environments may contain many individuals (and maybe many robots). So figuring out who is where, isolating an individual from a crowd, knowing whether one is being addressed, or timing an intervention in an ongoing conversation all present major difficulties that require beyond state-of-the-art solutions.
Even if some of these practical problems could be overcome, there are still issues concerning the role of *language* in human-robot interaction. For example, studies into the usage of smart assistants suggest that, far from engaging in a promised natural 'conversational' interaction, users tend to resort to formulaic language and focus on a handful of niche applications which work for them (Moore et al., 2016). Given the pace of technological development, it might be expected that the capabilities of such devices will improve steadily, but according to Phillips (2006) there is a 'habitability gap' in which usability drops as flexibility increases - see Figure 2.

[7]One well known robot even has its microphones mounted immediately adjacent to its cooling fans!

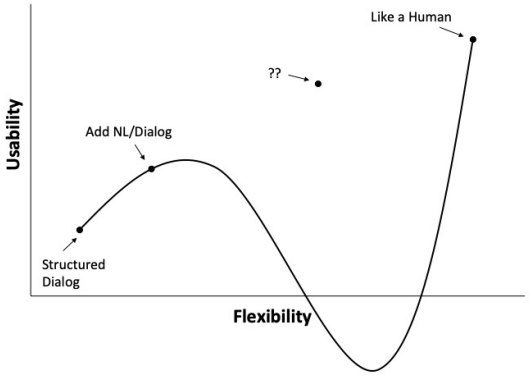


Figure 2: Illustration of the drop in usability that can occur in a spoken language dialogue system when its flexibility is increased.

It has been hypothesised that the habitability gap is a manifestation of the 'uncanny valley' effect (see Figure 3) whereby a near human-looking artefact (such as a humanoid robot) can trigger feelings of eeriness and repulsion (Mori, 1970). In particular, a Bayesian model of the uncanny valley effect (Moore, 2012) reveals that it can be caused by *misaligned* perceptual cues. Hence, a device with an *inappropriate* voice can create unnecessary confusion in a user. For example, the use of human-like voices for artificial devices encourages users to overestimate their linguistic and cognitive capabilities.

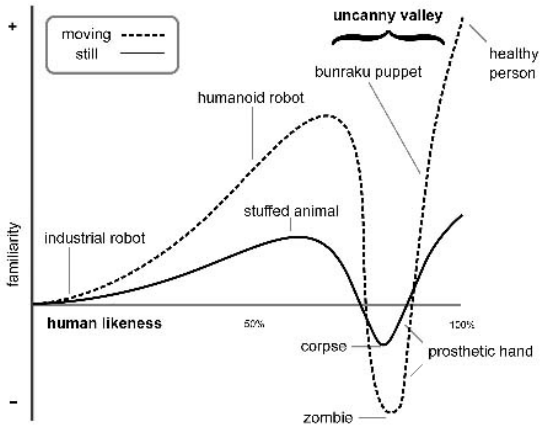


Figure 3: Illustration of the 'uncanny valley' effect in which a near human-looking artefact (such as a humanoid robot) can trigger feelings of eeriness and repulsion.

The Bayesian model of the uncanny valley effect suggests that the habitability gap can only be avoided if the visual, vocal, behavioural and cognitive *affordances* of an artefact are aligned. Given that the state-of-the-art in these areas varies significantly, this means that the capabilities of an artificial agent should be determined by the affordance with

the lowest capability (Moore, 2017; Wilson and Moore, 2017). In other words, emulating a human is a recipe for failure, rather "*it is better to be a good machine than a bad person*" (Balentine, 2007).

Another significant shortfall in our current level of knowledge about creating effective speech-based human-robot interaction is that robots need to *understand*, not just speak and listen. This is already a major impediment to conversational interaction with contemporary smart assistants. However, there is hope that deeper insights into the problem may arise from tackling language-based HRI on the basis that such interaction is necessarily *situated* and *grounded*; both of which are considered to be key aspects of genuine language understanding and give support to the 'pragmatics-first' view of language (Bar-On, 2017).

5.4. Ethical Issues

Finally, the drive towards speech-based human-robot interaction also raises a number of important ethical concerns. For example, the appearance of smart assistants in people's homes has already sparked controversy about whether such devices are listening to private conversations and sending sensitive personal information to unidentified third-parties. As a result, the level of *trust* that a user can place in an artificial conversational partner has become a subject of much debate.

Another area of concern is the ability to *fake* abilities that are far beyond the state-of-the-art. There are already examples of so-called 'intelligent' conversational robots being demonstrated to the public and the press which, on investigation, turned out to be operated by human beings, either remotely or even inside an elaborate robot costume! Such unethical activities tend to fuel the technological *hype* that often surrounds robots and speech-based interaction with them. Preprogrammed spoken responses to scripted verbal questions are easy to arrange, but at best seriously misrepresent the actual capabilities of the the device, and at worst undermines the confidence of funding agencies in determining what research (if any) needs to be supported.

6. Conclusion

Notwithstanding the tremendous progress that is currently taking place in spoken language technology, the achievement of effective speech-based human-robot interaction still raises a number of important challenges. Not only do the two fields of robotics and spoken language technology present their own special problems, but their combination raises an additional set of issues that are worthy of investigation. In particular, it is noted that there is a large gap between the type of formulaic speech-based interaction that typifies contemporary spoken language dialogue systems and the fully flexible natural language interaction exhibited in human-human conversation (Moore, 2016). Nevertheless, it is pointed out that the grounded and situated nature of speech-based human-robot interaction may lead to deeper insights into the pragmatics of language usage in real-world environments, thereby overcoming the current 'habitability gap'.

7. Bibliographical References

Balentine, B. (2007). *It's Better to Be a Good Machine Than a Bad Person: Speech Recognition and Other Exotic User Interfaces at the Twilight of the Jetsonian Age*. ICMI Press, Annapolis.

Bar-On, D. (2017). Communicative intentions, expressive communication, and origins of meaning. In Kristin Andrews et al., editors, *The Routledge Handbook of Philosophy of Animal Minds*, chapter 28. Routledge, London.

Boyd, C., (2018). *The past, present, and future of speech recognition technology*. https://medium.com/swlh/the-past-present-and-future-of-speech-recognition-technology-cf13c179aaf.

Huang, X. D. (2002). Making speech mainstream. Technical report, Microsoft Speech Technologies Group.

Moore, R. K., Li, H., and Liao, S.-H. (2016). Progress and prospects for spoken language technology: what ordinary people think. In *INTERSPEECH*, pages 3007–3011, San Francisco, CA. ISCA.

Moore, R. K. (2012). A Bayesian explanation of the 'Uncanny Valley' effect and related psychological phenomena. *Nature Scientific Reports*, 2(864):doi:10.1038/srep00864.

Moore, R. K. (2015). From talking and listening robots to intelligent communicative machines. In J Markowitz, editor, *Robots That Talk and Listen*, chapter 12, pages 317–335. De Gruyter, Boston, MA.

Moore, R. K. (2016). Is spoken language all-or-nothing? Implications for future speech-based human-machine interaction. In Kristiina Jokinen et al., editors, *Dialogues with Social Robots - Enablements, Analyses, and Evaluation*, pages 281–291. Springer Lecture Notes in Electrical Engineering (LNEE).

Moore, R. K. (2017). Appropriate voices for artefacts: some key insights. In *1st Int. Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots (VIHAR-2017)*, pages 7–11, Skovde, Sweden. VIHAR.

Mori, M. (1970). Bukimi no tani (the uncanny valley). *Energy*, 7:33–35.

Phillips, M. (2006). Applications of spoken language technology and systems. In Mazin Gilbert et al., editors, *IEEE/ACL Workshop on Spoken Language Technology (SLT)*, page 7, Aruba. IEEE.

Pieraccini, R. (2012). *The Voice in the Machine*. MIT Press, Cambridge, MA.

Smith, S., (2017). *Amazon Echo and Google Home to reside in over 50% of US households by 2022, as multi-assistant devices take off*. https://www.juniperresearch.com/press/press-releases/amazon-echo-google-home-to-reside.

Wilson, S. and Moore, R. K. (2017). Robot, alien and cartoon voices: implications for speech-enabled systems. In *1st Int. Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots (VIHAR-2017)*, pages 40–44, Skovde, Sweden. VIHAR.

Proceedings, pages 51–55
Paris, UNESCO Headquarters, 5-6 December, 2019.

Bangla Speech Synthesizer System for Bangladesh

Mohammad Nurul Huda, Sabbir Arif Siddique, Ishteaque Alam
Department of Computer Science and Engineering
United International University, Dhaka, Bangladesh
mnh@cse.uiu.ac.bd, sabb.a.sidd@gmail.com, ishteaque.ark@gmail.com

Abstract

This paper illustrates the design and implementation of Bangla (widely used as Bengali) Text to Speech (TTS) system from the very raw level without using any third party speech synthesis tool. For constructing the system we have considered two directions, where one is based on phoneme and another one is on syllable. In this study, our proposed system comprises some stages. At first stage audio sounds are recorded for each of the Bangla phonemes and three thousand out of 250000 syllables in Bangla, and then noise is reduced to obtained high quality sounds for each phoneme and syllable. Second stage searches for longest possible matching of the syllables if it is available in the input text, and if not, then searches for the phonemes to match with the corresponding graphemes. For further improvement, we also added the complex conjuncts which need to be handled separately. It is observed from the experiments that the syllable based method provides the better quality speech for the input text in comparison with the method based on phoneme.

Keywords: ***text to speech; speech synthesis; phoneme; syllable; graphemes***

Résumé

এই গবেষণাটি বাংলা লেখ্য ভাষাকে কথ্য ভাষায় রূপান্তর করে। এই কাজটি দুইভাবে করা যায়, একটি হলো ধ্বনির মাধ্যমে অন্যটি হলো শব্দাংশ এর মাধ্যমে। এই গবেষণাটিতে, আমাদের প্রস্তাবিত সিস্টেমটি কয়েকটি স্তর নিয়ে গঠিত। প্রথম পর্যায়ে প্রতিটি বাংলা ধ্বনির জন্য অডিও শব্দের রেকর্ড করা হয় এবং বাংলায় ২৫০০০০ শব্দাংশের জন্য তিন হাজার শব্দাংশ রেকর্ড করা হয় এবং তারপরে প্রতিটি ধ্বনির এবং শব্দাংশের জন্য উচ্চমানের শব্দ পাওয়ার জন্য নয়েস কমিয়ে আনা হয়। দ্বিতীয় পর্যায়ে শব্দাংশের দীর্ঘতম সম্ভাব্য মিল খুঁজে পাওয়ার চেষ্টা করা হয় যদি এটি ইনপুট টেক্সট এ পাওয়া যায় এবং যদি না পাওয়া যায়, তবে ধ্বনিযুক্ত শব্দগুলিকে অনুরূপ গ্রাফিম এর সাথে মিল রেখে অনুসন্ধান করা হয়। অধিক উন্নত সিস্টেম পাওয়ার জন্য, আমাদের যুক্তাক্ষর গুলোকে আলাদাভাবে চিন্তা করা দরকার। পরীক্ষাগুলি থেকে এটি পর্যবেক্ষণ করা হয়েছে যে শব্দাংশ ভিত্তিক পদ্ধতি ধ্বনি ভিত্তিক পদ্ধতির তুলনায় ভাল মানের ফলাফল প্রদান করে।

1. INTRODUCTION

Speech synthesis is the automatic production of human speech, where a computer system used for this purpose is called a speech synthesizer, which can be implemented in software or hardware products, but a text to speech (TTS) system converts normal langauge text into speech [1]. Mute people cannot talk, but they will be able to talk using the TTS system, where they will type their desired words or sentences, and the TTS system will convert them into human speech. On the other hand, blind people cannot see, but they can hear the sound. If a mute want to communicate with a blind, the blind cannot see mute's sign language, but he/she can hear the speech that has been produced by the TTS system. Therefore, a mute can converse with a blind using the TTS system [2, 3].

Several attempts [4-8] had been made to develop and stimulate the process of development of the Bangla TTS synthesis system. In [4], epoch synchronous non overlap add (ESNOLA) method based concatenative speech synthesis system for Bangla was developed by Shyamal Kumar Das Mandal, et. al, in which authors described a system for concatenative speech synthesis using ESNOLA technique. Again, S. K. D. Mandal, et. al [5] showed some practical applications of Bangla TTS system using Epoch Synchronous Non Overlap Add (ESNOLA) technique. On the other hand, some important aspects of Bengali Speech Synthesis System proposed by A. Bandyopadhyay [6] used phonemes to develop voice database and used Epoch Synchronous Overlap Add (ESOLA) technique to concatenate the phonemes. Besides, T. Sarkar, et. al [7] described about grapheme to phoneme conversion, optimal text selection, automatic segmentation tools and shown their experiment results. Moreover, Firoj Alam, et al [8] described the development process of Bangla (widely used as Bengali) TTS using a speech synthesis tool named Festival. But very few litteratures are found in Bangla spoken by Bangladeshi people.

In this paper, we have proposed a system that shows the design and implementation of Bangla Text to Speech (TTS) system from the very raw level without using any third party speech synthesis tool. Two proposed systems in this study based on phonemes and syllables comprises two stages, in which the first stage audio sounds are recorded for each of the Bangla phonemes and three thousand out of 250000 syllables in Bangla, and then noise is reduced to obtained high quality sounds for each phoneme and syllable; and the second stage searches for longest possible matching of the syllables if it is available in the input text, and if not, then searches for the phonemes to match with the corresponding graphemes. For further improvement, we also added the complex conjuncts which need to be handled separately.

2. SPEECH SYNTHESIS

Speech synthesis is the computer-generated simulation of human speech, which is used to translate written information into aural information where it is more convenient [9]. The generation of a sound waveform of human speech from a t extual or phonetic description is called speech synthesis [10]. To generate speech output from a given text, first, the input text is analyzed deeply. Then grapheme to phoneme conversion is carried out using pronunciation and letter to sound rule. Same phoneme or syllable may have different pronunciations depending on the grammatical and pronunciation rules. So the

pronunciation of the phonemes and syllables are detected by analyzing those rules. After identifying the corresponding sounds of matched syllable and phonemes, they are concatenated and played to generate expected speech output. Fig. 1 shows the workflow of speech synthesis.

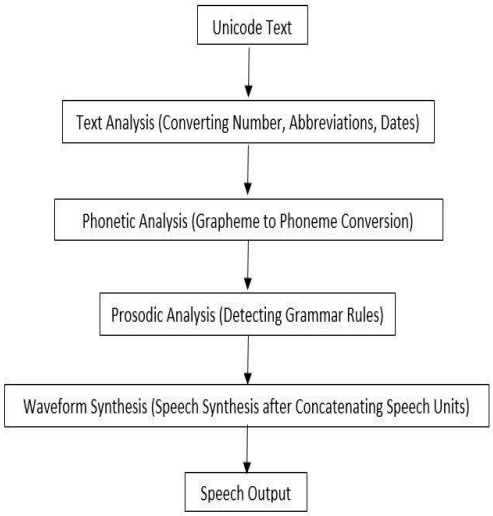


Fig. 1. Work flow of Speech Synthesis.

3. SPEECH SYNTHESIS METHODOLOGY

Different phases of our speech synthesis methodology are described below.

3.1 Text Analysis

In this level, the input text is analyzed deeply to convert it into pronounceable sounds. Bangla text contains the following alphabets and symbols. Here, IPA symbol for phoneme is also shown here

Bangla Vowel (বাংলা স্বরবর্ণ)

অ	আ	ই	ঈ	উ	ঊ	ঋ	এ	ঐ	ও	ঔ
a	ā	i	ī	u	ū	ṛ	e	ai	o	au
[ɔ, o]	[ɑ:]	[i, e]	[i]	[u, o]	[u]	[ri]	[ɛ, æ]	[oi]	[o]	[ow]
ক	কা	কি	কী	কু	কূ	কৃ	কে	কৈ	কো	কৌ
ka	kā	ki	kī	ku	kū	kṛ	ke	kai	ko	kau

Bangla Consonants (বাংলা ব্যাঞ্জনবর্ণ):

ক	ka	[kɔ]	খ	kha	[kʰɔ]	গ	ga	[gɔ]	ঘ	gha	[gʱɔ]	ঙ	ṅa	[ŋɔ]
চ	ca	[ʧɔ]	ছ	cha	[ʧʰɔ]	জ	ja	[ʤɔ]	ঝ	jha	[ʤʱɔ]	ঞ	ña	[nɔ]
ট	ta	[ʈɔ]	ঠ	tha	[ʈʰɔ]	ড	da	[ɖɔ]	ঢ	dha	[ɖʱɔ]	ণ	ṇa	[nɔ]
ত	ta	[tɔ]	থ	tha	[tʰɔ]	দ	da	[dɔ]	ধ	dha	[dʱɔ]	ন	na	[nɔ]
প	pa	[pɔ]	ফ	pha	[pʰɔ]	ব	ba	[bɔ]	ভ	bha	[bʱɔ]	ম	ma	[mɔ]
য	ya	[ʤɔ]	র	ra	[rɔ]	ল	la	[lɔ]						
শ	śa	[ʃɔ/sɔ]	ষ	ṣa	[ʃɔ]	স	sa	[ʃɔ/sɔ]	হ	ha	[ɦɔ]			
য়	ya	[jɔ]	ড়	ṛa	[ɽɔ]	ঢ়	ṛha	[ɽʱɔ]						

Modifier Symbols:

্	hasanta - mutes inherent vowel	ক্	k [k]
ৎ	khanda-ta - final unaspirated dental	কৎ	Kat [kɔt̪]
ং	anusvāra - final velar nasal	কং	kaṁ [kɔŋ]
ঃ	visarga - adds voiceless breath after vowel	কঃ	kaḥ [kɔh] / [kɔ]
ঁ	chandra-bindu - nasalises vowels	কঁ	km̐ [kɔ̃]

Post-consonantal vowel signs:

আ = া (akar) as in কা (should be after consonant)

ই = ি (hrossikar/ikar) as in কি

ঈ = ী (dirgghikar/ikar) as in কী (should be after consonant)

উ = ু (hrossukar/ukar) as in কু (should be under consonant)

ঊ = ূ (dirgghukar/ukar) as in কূ (should be under consonant)

ঋ = ৃ (rikar) as in কৃ (should be under consonant)

এ = ে (ekar) as in কে

ঐ = ৈ (oikar) as in কৈ

ও = ো (okar) as in কো

ঔ = ৌ (oukar) as in কৌ

First, total number of characters in input text is counted. From the beginning of the input text character, longest possible match for the synthesized sound unit is searched. Words are differentiated with space so the search continues to find longest sub-word match until a space is found. If search pointer reaches to a space, new word begins. The search continues until the end of the input text.

To illustrate the analyzing process let us take a sample text and analyze it:

Sample Bangla text: "আমি ভাত খাই"

Here, total number of characters = 11

Total spaces = 2

Total pronounceable characters = 11 - 2 = 9

Total words = 2

আমি = আ + ম +

ভাত = ভ + া +ত

খাই = খ + া + ই

The Non Standard Words (NSW) need to be normalized. Non Standard Words include abbreviations, acronyms, currency, dates, numbers (year, time, ordinal, cardinal, floating point). These Non Standard Words needs to be converted to standard words or syllables.

Example of Non Standard Words (NSW):

Example of currency: ১০০০/- = এক হাজার টাকা

Example of date: ২২-৪-২০১৪ = বাইশ চার দুই হাজার চৌদ্দ

Example of number: ১ = এক, ২ = দুই, ৩ = তিন, ৪ = চার, ৫ = পাঁচ, ৬ = ছয়, ৭ = সাত, ৮ = আট, ৯ = নয়, ০ = শুন্য

Example of time: ১২:৪০ = বারোটা চল্লিশ

There are several conjuncts in Bangla language. Some of them are shown below:

Bangla Conjuncts (বাংলা যুক্তবর্ণ):[3.2]

ক্ক = ক + ক; Example- আক্কেল, টেক্কা

ক্ট = ক + ট; Example- ডক্টর (Comment: Basically used in English/foreign debt words)

ক্ট্র = ক + ট + র; Example- অক্ট্রয়

ক্ত = ক + ত; Example- রক্ত

3.2 Phonetic Analysis

The method of finding pronunciation of input text is analyzed in this level. In our TTS system, we have used the following techniques for phonetic analysis:

3.2.1. Phoneme based technique for phonetic analysis: At first, we used the phonemes of Bangla language in our system. We have recorded the phonemes and implemented in our system by grapheme to phoneme conversion to generate speech output, but the result was not satisfactory. Let us give an example of the technique:

The Bangla word "তোমার" will be pronounced as following, if the phonemes are used only:

তোমার = ত্ + ও + ম্ + আ + র্

which is: /t/ + /o/ + /m/ + /a/ + /ɣ/

This pronunciation is not good enough to understand. So, we decided to think in another direction. Then we found that if we can use the syllables, the output will be more satisfactory. So, then we started working with the syllables.

3.2.2. Syllable based technique for phonetic analysis:

We started recording the syllables and implemented them in our system by searching longest possible match with the syllables from the beginning of the words. But, later on we observed that longest possible match is not the best option always. So, we started analyzing the syllables and implemented them after finding and sorting them in the order which suits the best combination for pronunciation. To illustrate the syllable analyzing process, let us take a sample text and analyze it to find the best combination of syllable for pronunciation :

Sample Bangla Text: তোমার - তোমাদের - তোমরা

Here, the word "তোমার" can be a good example for analyzing Bangla text. If we select "তোমা" as a syllable and "র" as another syllable as we have searched for the longest possible match from the beginning of the input text first, the word "তোমার" will be pronounced as following:

তোমা + র্

This pronunciation is not good enough for "তোমার" but this "তোমা" will be good for pronunciation of "তোমাদের"

If we select "তোম" as a syllable and "ার" as another syllable, the word "তোমার" will be pronounced as:

তোম্ + আর্

This pronunciation is also not good for "তোমার" but this "তোম" will be good for pronunciation of "তোমরা" .

If we select "তো" as a syllable and "মার" as another syllable, the word "তোমার" will be pronounced as:

তো + মার্

This pronunciation is good for "তোমার" and this "তো" will be also good for "তোমাদের".

So, তো + মার is the best combination to make pronounceable "তোমার".

Thus, the analysis process is done for "তোমার - তোমাদের - তোমরা".

3) Syllable and Phoneme based technique for phonetic analysis: While working with the syllables, we have come to know that there are more than 200000 syllables in Banga language. The more syllables we can used, the more the performance will increase. But, it was not possible to work with almost 250000 thousand syllables. So, we used the most common three thousand syllables and phonemes together according to our requirement.
First, the system will search for longest possible match of the syllables. If there is no such syllable found, then it will search for the phonemes to match with the corresponding graphemes. We observed that the performance was getting better.

Let us take the following example to illustrate the syllable + phoneme technique:

Suppose we have the syllables "তো", "মার", and "দের" in our voice database as we have seen in the previous technique that this is the best combination of syllables to pronounce "তোমার" and "তোমাদের". And suppose, we do not have the syllable "তোম্", but we have "তো" and "রা" in our database.

Yet the word "তোমরা" can be pronounced as the following as we have used the phonemes with the syllables:

তোমরা = তো + ম্ + রা। So, in this technique, we need all the phonemes and the basic common syllables of a language to develop its TTS system. By this technique, we can cover the whole language in our system with better pronunciation.

3.3 Prosodic Analysis

Same phoneme or syllable may have different pronunciations depending on its prior and post characters or even its position in the word. The pronunciations will be according to the grammatical and pronunciation rules of Bangla language. To apply those rules, prosodic analysis is required. This prosodic analysis process is discussed below using examples.

Let us take a look at the following simplest Bangla pronunciation rules to understand the prosodic analysis process:

Bangla Pronunciation Rule #1 (a):*

If a consonant letter appears in the beginning of a Bangla word and if the post character of this consonant is a consonant letter, the first consonant will be pronounced as[11]:

Phoneme of the consonant + "অ"

Bangla Pronunciation Rule #1 (b):*

And if a consonant letter appears in the beginning of a word and if the post character of this consonant is a vowel, then the consonant will be pronounced as:

Phoneme of the consonant + "ব"

For example, let us take the consonant ÓeÓ.

The word "বক" has got first character "ব" and its post character is "ক" which is a consonant. So, according to Bangla pronunciation rule #1 (a), the word "বক" will be pronounced as:

বক = ব্ + অ + ক্

In the word "বই", the first character is also "ব" but its post character is "ই" which is a vowel. So, according to Bangla pronunciation rule #1 (b), the word "বই" will be pronounced as:

বই = ব্ + ও + ই

Similarly, in the word "বল", the first character is "ব" and its post character is "ল" which is a consonant. So, according to Bangla pronunciation rule #1 (a), the word "বল" will be pronounced as:বল = ব্ + অ + ল্

And in the word "বউ", the first character is also "ব" but its post character is "উ" which is a vowel. So, according to Bangla pronunciation rule #1 (b), the word "বউ" will be pronounced as:

বউ = ব্ + ও + উ

In the same way,

মগ = ম্ + অ + গ্ but, মই = ম্ + ও + ই

কম = ক্ + অ + ম্ but, কই = ক্ + ও + ই

3.4 Waveform Synthesis

The waveform is synthesized step by step. The steps are given in Fig. 2.

1) Record Sound: The sound is first recorded for corresponding phonemes and syllables. *Fig. 3 shows recorded sound wave with noise.* The recording should be in a sound proof place, otherwise there will be so much noise and interference in the recording, which cannot be reduced.

2) Convert Sound to Wave Signal: After recording the sounds, they are converted to wave signals. The wave signal of a recorded sound is shown below.

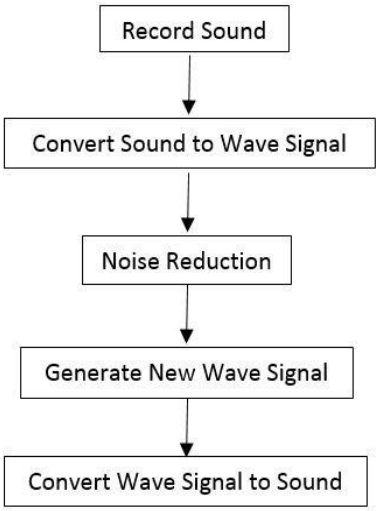


Fig. 2. Waveform Synthesis

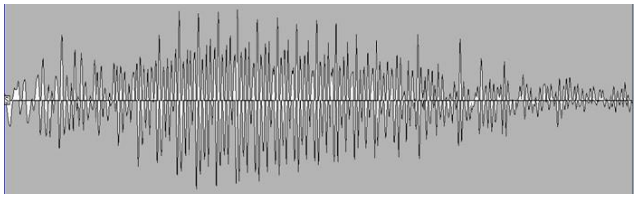


Fig. 3. Wave Signal of a Recorded Sound (with noise)

3) Noise Reduction: The noise in the wave signal is cut down to reduce the noise of the sound. The noise of the above wave signal is cut down to reduce the noise of the sound.

4) Wave signal by reducing the noise: After reducing the noise from the wave signal, new wave signal is generated *and is shown in Fig. 4.* The new noise free wave signal of the above wave signal is shown below.

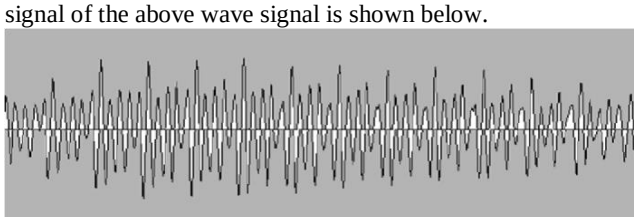


Fig. 4. Wave Signal after Noise Reduction

5) Convert Wave Signal to Sound: After generating the new noise free wave signal, the signal is converted to sound. This sound is the required noise free sound which can be used to generate speech output of a given text.

Concatenate Sound Units to Generate Output: The noise free sound units are concatenated by the application according to the given input text to generate the desired speech output.

4. EXPERIMENTS

4.1 Corpus

We have prepared the voice database by renaming the audio files by their corresponding syllable or phoneme names by which we can search easily to find and concatenate them to generate the desired speech output. O

4.2 Experimental Setup

The application has two methods to take input text. User can write on the given textbox or can open a text file to be read. Texts in the above picture are inputted by opening a text file tested for experiment. The output was good enough to be understood.

5. EXPERIMENTAL RESULT ANALYSIS

As we worked with the phonemes first, then the syllables, we got different results for phonemes and syllables. The result of the experiments with phonemes and syllables are shown below with a sample text. Suppose, the database has the following phonemes:

ত্ , ও, ম্, আ, র্, দ্, এ

and the following syllables:

তো, মার, মা, দের, তোম্

So, with these phonemes and syllables the result with the sample text will be shown in Table 1.

As we told earlier that Bangla language has huge number of syllables, so we could not include them all in our system, but we found a way- that is use all phonemes and most common syllables together. Now, suppose our database does not have the syllable "তোম্", but has all the phonemes. So the result will be shown in Table 2. So the result of the different phonetic analysis technique can be compared and shown in Table 3.

TABLE I. EXPERIM ENTAL RESU LTS FOR PHONEMES AND SYLLABLES

Sample text: "তোমার - তোমাদের - তোমরা"			
Pronunciation	তোমার	তোমাদের	তোমরা
Using Syllables	তো + মার	তো + মা + দের	তোম্ + রা
Using Syllables + Phonemes	তো + মার	তো + মা + দের	তো + ম্ + রা

TABLE II. EXPERIM ENTAL RESU LTS FOR PHONEME + SYLLABLES

Sample text: "তোমার - তোমাদের - তোমরা"			
Pronunciation	তোমার	তোমাদের	তোমরা
Using Phonemes	ত্ ম্ + ও + র্ + আ +	ত্ + ম্ + ও + আ দ্ + এ + র্	ত্ ম্ + ও + র্ + আ
Using Syllables	তো + মার	তো + মা + দের	তোম্ + রা

TABLE III. COM PARISON O F DIFFERENT TECHNIQU ES USIN G PHONETIC ANA LYSIS

Using	Pronunciation	Coverage of input text
Phonemes only	Cannot be understood properly	Covers All input text
Syllables only	Good enough to be understood	Do not cover all input text
Syllables + Phonemes	Good enough to be understood	Covers All input text

6. CONCLUSION

This paper has showed a technique for Bangla text to speech and concludes the following:

i) This research was done from the very raw level, starting from using our own voice recordings to create phonemes and syllables. ii) More than 3000 syllables and phonemes were used during the development process. iii) Syllable based method showed high quality speech than the phoneme based method

In near future the author would like to do synthesis by covering the whole Bangla grammar and doing text normalization for larger context. The author would also work on conjuncts and more syllables in future. Besides, the experiments for the existing system to compare with our proposed method are not presented here. The author would like to these experiments for the future study.

REFERENCE

[1] Speech Synthesis
Website:http://en.wikipedia.org/wiki/Speech_synthesis

[2] Crowdsourcing helps Bangladesh's blind pupils
Website:http://www.dw.de/crowdsourcing-helps-bangladeshs-blindpupils/a-17744891

[3] Bangla text to Speech using Festival by Firoj Alam
Website:
https://www.academia.edu/2955759/Bangla_Text_to_Speech_using_Fes tival

[4] Shyamal Kumar Das Mandal and Asoke Kumar Datta,"Epoch Synchronous Non Overlap Add (ESNOLA) Method based Concatenative Speech Synthesis System for Bangla," Centre for development of Advanced Computing (C-DAC), Kolkata, India

[5] S.K.D. Mandal and B. Pal, "Bengali Text to Speech Synthesis System: A Novel Approach for Crossing Literacy Barrier," CSI-YITPA(E), 2002

[6] A. Bandyopadhyay, "Some Important Aspects of Bengali Speech Synthesis System," IEMCT, 2002.

[7] T. Sarkar, V. Keri, M. Santhosh and K. Prahallad, "Building Bengali
Voice Using Festvox, " CLSI 2005

[8] The Festival Speech Synthesis System
Website:http://www.cstr.ed.ac.uk/projects/festival/

[9] Speech Synthesis
Website:http://whatis.techtarget.com/definition/speech-synthesis
Speech Synthesis
http://dictionary.reference.com/browse/speech+synthesis

Proceedings, pages 56–60
Paris, UNESCO Headquarters, 5-6 December, 2019.

A Real Time Instruction Extractor from Traffic Signal for Translation

Professor Dr. Mohammad Nurul Huda, Sabbir Arif Siddique
United International University, eGeneration Limited
mnh@cse.uiu.ac.bd, sabb.a.sidd@gmail.com

Abstract

This paper has developed and demonstrated a system to build traffic instruction detection and translation tools that can extract and convert Bangla text from natural images containing traffic instruction. In the process of developing the system, we have applied various techniques to extract and convert information from natural images. These techniques involve Image Processing, Machine Learning, Optical Character Recognition and Machine Translation. The proposed system consists of three steps, which are Text extraction from image, Post Processing by Language Model and Machine Translation.

Keywords: Optical Character Recognition, Image Processing, Machine Translation, Language Model

Résumé

এই গবেষণাটি দ্বারা আমরা একটি মেশিন লারনিং সমৃদ্ধ ট্রাফিক সিগন্যাল অনুবাদক টুল তৈরি করেছি। এই টুলটি বাংলা অপ্টিকাল ক্যারেক্টার রেকগনিশন ও মেশিন ট্রান্সলেশন এর সম্মিলনে তৈরি হয়েছে। এই টুলটি ব্যাবহার করে বিদেশী ব্যাক্তি বাংলা ট্রাফিক সিগন্যাল ইংরেজি ভাষায় অনুবাদ করে পরতে ও বুঝতে পারবে।

1. Paper Submission

The problem of understanding traffic signs in Bangla has been identified as a major problem for the foreigners. As these traffic signs contain both images of visual traffic signal along with Bangla text, it is nearly impossible to acknowledge the signs for a foreign citizen. Figure 1 is an illustration of a few existing traffic signs found on the roads of Dhaka.

Figure 1: Traffic Signs in Bangladesh.

Moreover, placement of traffic signs does not follow any international standard. Therefore, it may be rather difficult for non-local residents to find the signs without much effort. In our paper, we have proposed a state-of-the-art solution to address the mentioned problems. In this study we have used image processing mechanism and machine translation in this purpose.

The main goal of our image processing part of this research is to analyze a captured image, find, and segment the Bangla letters from there. In addition, we have also incorporated an efficient machine translator to translate the extracted Bangla text into English and other major languages.

The paper [1] proposed a novel system for the automatic detection and recognition of text in traffic signs. The authors have proposed a system in their work, which is capable of defining search area within the image. The paper [2] has recommended a system that can detect and recognize instruction from traffic signals. The authors of this paper have proposed a system to integrate in the Advanced Driver Assistance System (ADAS). We have recognized and implemented techniques expressed and illustrated in this paper. Moreover, we have incorporated additional techniques to improve the outcome of the Bangla OCR. These techniques include Edge Detection using Canny method [3], Gaussian filter [4], Edge Tracking by Hysteresis [5], B/W labeling, Character Segmentation [6], Character Recognition through Back Propagation Neural Network [7] to process the text extracted from the image and Example Based Machine Translation [8] algorithm.

Our proposed method for Bangla detection and translation from traffic signs is comprised of three stages. The first stage detects the traffic signs from natural images. In consequence, the second stage extracts Bangla text from the natural image. In the final stage, the text is translated into English. This paper represents the first endeavor in developing a traffic sign detection and translation system for Bangla language. Although Google and Bing have similar products, they however do not have support for Bangla yet.

2. Previous Study

There are many works for Bangla OCR from documents like Bangla OCR by UIU and first commercial OCR “Puthi OCR” by Team Engine. Most prominently there are two notable thesis work for Bangla OCR from image. The first one is from Khulna University by Zahid et.al and other one is from Computer Vision & Pattern Recognition Unit,

Indian Stat. Inst., Kolkata, India. In this research, we have incorporated techniques analyzed from the above-mentioned sources and combined them into a single system application.

3. Proposed System

The proposed system processes the captured images and converts them into English instructions. Distinct modules of the system execute in sequence to acquire the targeted goal from the input. Each of these modules employs diversified tools and contemporary algorithms. These modules are explained with demonstration and relevant diagrams in the following section. The proposed system is illustrated in the system diagram in Figure 2.

3.1 Image Processing

Captured image that contains Bangla traffic instruction is processed through a sequence of techniques, which are clarified by demonstration in the following sub sections.

3.1.1 Pre-Processing

After the natural image is captured, the preprocessing mechanism is conducted on the image. The input and output of the process is illustrated in Figure 3.

Preprocessing resizes and adjusts the RGB value of the captured images. The outcome of this stage is the B/W image with the corrected proportion.

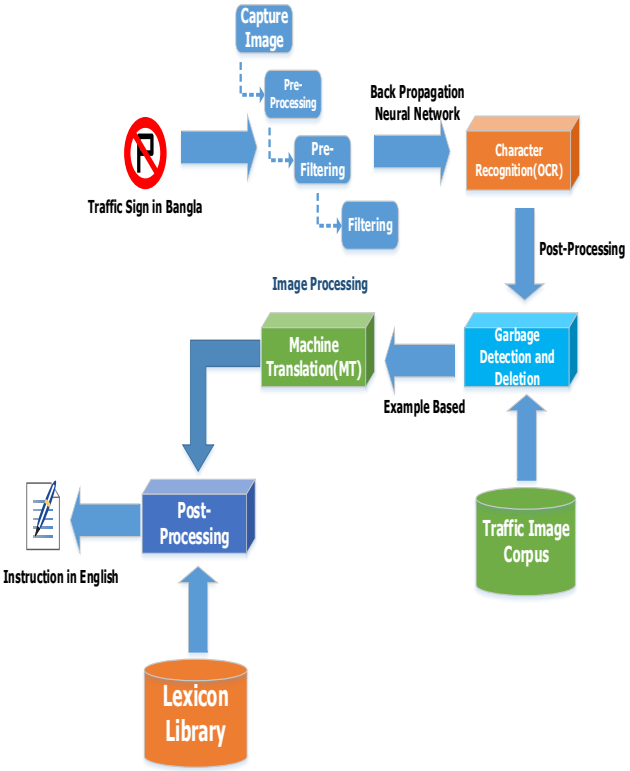


Figure 2: Proposed System Diagram.

Figure 3: Capturing and Preprocessing of Natural Image.

3.1.2 Pre-Filtering

On completion of the preprocessing pre-filtering is applied

on the processed image. The pre-filtering of the image is conducted by employing Edge Detection by Canny Edge Detection Method. The Canny Method is less likely than other methods to be fooled by noise. The general criteria for edge detection include the following steps.

I. Detection of edge with low error rate, which means that the detection should accurately catch as many edges shown in the image as possible.

II. The edge point detected from the operator should accurately localize on the center of the edge.

III. A given edge in the image should only be marked once, and where possible, image noise should not create false edges.

After edge detection process is conducted, Gaussian filter [4] is applied on the output to further fine-tune the detected edges. The equation for a Gaussian filter kernel with the size of (2k+1) * (2k+1) is shown as following:

$$H_{ij} = \frac{1}{2\pi\sigma^2} * \exp(-\frac{(i-k-1)^2+(j-k-1)^2}{2\sigma^2})$$

Here is an example of a 5x5 Gaussian filter, used to create the image to the right, with = 1.4. Here the asterisk denotes a convolution operation.

$$\mathbf{B} = \frac{1}{159}\begin{bmatrix} 2 & 4 & 5 & 4 & 2 \\ 4 & 9 & 12 & 9 & 4 \\ 5 & 12 & 15 & 12 & 5 \\ 4 & 9 & 12 & 9 & 4 \\ 2 & 4 & 5 & 4 & 2 \end{bmatrix} * \mathbf{A}.$$

After applying the filter, the intensity gradient of the image is established. The edge detection operator (Roberts, Prewitt, and Sobel for example) returns a value for the first derivative in the horizontal direction (Gx) and the vertical direction (Gy). From this, the edge gradient and direction can be determined by the following equations.

$$\mathbf{G} = \sqrt{\mathbf{G}_x^2 + \mathbf{G}_y^2}$$

$$\mathbf{\Theta} = \operatorname{atan2}(\mathbf{G}_y, \mathbf{G}_x)$$

In consequence, edge-thinning technique termed Non-maximum suppression is enforced on the produced output. After application of non-maximum compression, the edge pixels are quite accurate to present the real edge. However, there are still some edge pixels at this point caused by noise and color variation. In order to get rid of the spurious

responses from these bothering factors, it is essential to filter out the edge pixel with the weak gradient value and preserve the edge with the high gradient value. Thus, two threshold values are set to clarify the different types of edge pixels, one is called high threshold value and the other is called the low threshold value. On resolving the double threshold value, edge tracking is conducted by Hysteresis. Afterwards structural elements of the image are extracted and then dilated. On the completion of the above-mentioned processes, the cropped images are acquired. The cropped elements along with some garbage is illustrated in

Figure 4.

Figure 4: Cropped Elements after Pre-Filtering.

3.1.3 Filtering

Filtering techniques is further applied on the pre-filtered output. These techniques include range estimation of the pre-filtered output. In the mentioned process, the garbage elements are removed and actual Bangla texts from the image is revealed. The flow diagram of the filtering process is illustrated Figure 5.

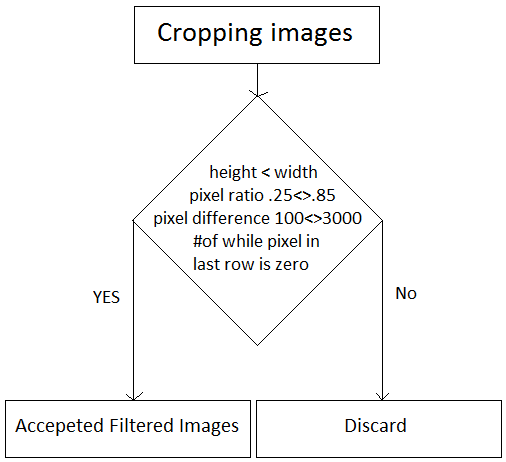


Figure 5: Flow Diagram of Filtering.

The output of this stage is illustrated in Figure 6.

Figure 6: Filtered Image.

3.1.4 Character Segmentation

The character segmentation process segments the characters in two categories. The first category is Characters without KAR (Bangla −□□□).

Figure 7: Characters without KAR.

The second category is characters with KAR.

Figure 8: Characters with KAR.

Therefore, the final output will be like the illustration in Figure 9.

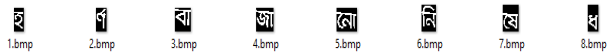


Figure 9: Final Output of Character Segmentation.

As we train our NN with black letters which have white background so after segmenting those letters we simply reverse the black pixel with white pixels of each letters and the output is given in Figure 10.

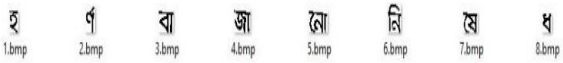


Figure 10: Output after Pixel Reversing.

In addition, for better image processing we reshape the image into a constant height and width. We use 45 × 45 (= 2025) constant shapes for each letter. This output is illustrated in Figure 11.

Figure 11: Characters after Reshaping.

The flow diagram of the character segmentation is illustrated below in Figure 12.

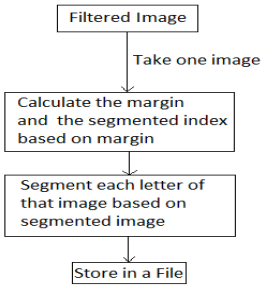


Figure 12: Flow Diagram of Character Segmentation.

3.2 Character Recognition and Post-Processing

After segmentation process, the output must be converted into machine-readable text. Neural network is employed to generate that conversion. However, the output of the neural net may contain a few garbage, which must be eliminated

to extract clean text. The processes in Figure 12 are detailed in the following sub-sections.

3.2.1 Character Recognition Using BP ANN

Backpropagation Artificial Neural Network (BP ANN) is employed in the proposed system to convert the segmented characters into electronic text. The text is retrieved in Unicode font. Backpropagation (BP) artificial neural network is the most commonly used algorithm in OCR, as it is highly effective in the given context. A typical BP ANN is illustrated in Figure 13.

BP ANN employs the following technique to extract the electronic character from the character segmentation output, which is depicted in Figure 14.

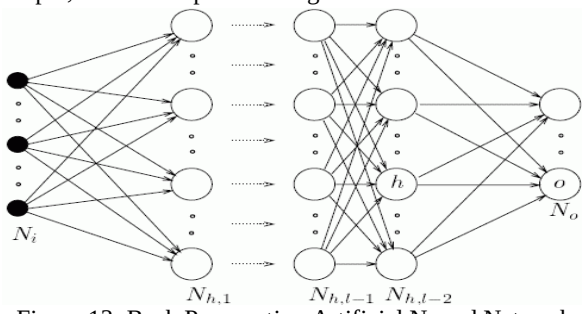


Figure 13: Back Propagation Artificial Neural Network.

3.2.2 Garbage Detection and Deletion

After character segmentation, post processing is conducted. Post processing is primarily consisted of garbage detection and deletion. To detect garbage from multi characters we will perform a partial string matching. Partial string matching is an approach to identify garbage value and useful to predict words from a partially correct word. Therefore, here is our algorithm.

- Split the result string.
- Iterate through all words.
- if(word.length > 1)
  Perform partial matching for each of the Bangla words in dictionary. Find the best matched Bangla words and return.

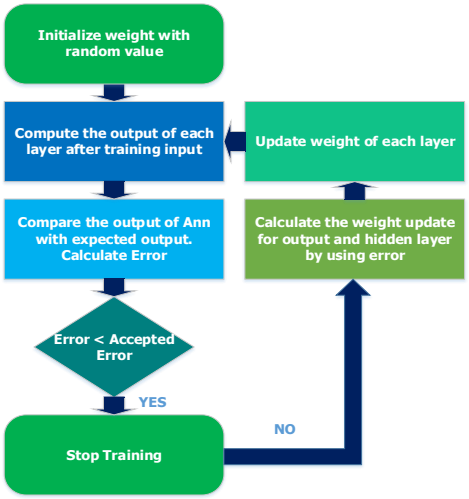


Figure 14: Flow Diagram of Character Recognition

Afterwards Levenshtein's Distance is employed to acquire best matching strings from the string dictionary. In figure 15 there is an illustration of an extracted charaqcter string with garbage values.

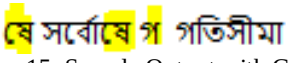


Figure 15: Sample Output with Garbage.

Now the first ষে in Figure 15 will be removed as that will not be partially matched with any word. গ in Figure 15 will also be removed, as it is a single character.

Instead of detecting সর্বোচ্চ our BP ANN returns সর্বোষে but it's partially matched so that will be replaced with correct one! Partially matched with সর্বোচ্চ because it will need three moves to transform one to another, which is minimum among other words in dictionary, and similarity between two words is 62%. Hence after post processing of the sample output we acquire the clean and authentic string as illustrated in Figure 16.

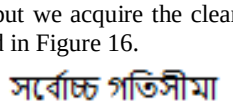


Figure 16: Output after Post Processing.

3.3 Machine Translation [8]

Now we have successfully extracted the authentic and clean text from the natural image. The next step is to convert the Bangla text into English. Machine Translation is a process of translating one word/sentence to another language's corresponding word/sentence. Machine translation is a complex problem because there are thousands of things that are needed to be considered. In basic level, we can just replace the words in a sentence with corresponding word in target language. That is not able to produce a good translation as the sentence structures are different and the recognition of whole phrases with their closest counterparts in the target language is needed. The approach [8] that we have taken in this paper is illustrated in Figure 17.

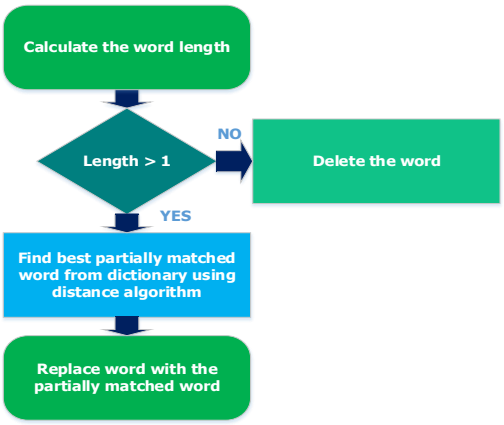


Figure 17: Flow Diagram of Garbage Deletion Process.

This approach can successfully translate most of the common traffic instructions. However, the English meanings that are constructed using multiple Bangla words is not considered here. The process of our approach of machine translation is illustrated in the Figure 18. The translated output of the extracted Bangla text is shown in Figure 19.

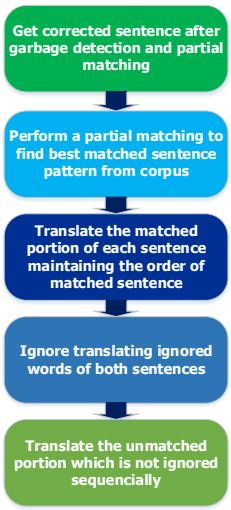


Figure 18: Flow Diagram of Machine Translation.

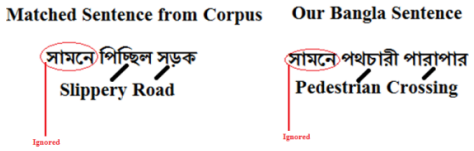


Figure 19: Machine Translation of Extracted Bangla Text.

4. Experimental Result and Analysis

The image of traffic instructions is very rare in internet. In fact, the traffic instructions are hard to find. Therefore, we really did not manage to get a plenty number of images for training and testing. That is one of the biggest difficulties that we have faced. Therefore, we have to test our system for limited training and testing data.Training Image Corpus:22 and test corpus:6. Backpropagation Artificial Neural Network is used where input features: 45×45 and output size : 18. Number of words in Traffic Instruction Database: 45. A demo corpus of pattern matching is shown in Table 1 and experimental result is shown in Table 2.

Table 1: Demo Corpus for Pattern Matching.

Bangla Sentence	English Sentence
পার্কিং নিষেধ	No parking
পথচারী চলাচল নিষেধ	No Pedestrian
সামনে টি-জাংশন আছে	T-junction
সামনে পথচারী পারাপার	Pedestrian Crossing

5. Conclusion and Future Works

In this research work, state of the art algorithms to translate Bangla Traffic sign into English for Foreigners were implemented. Because Canny edge detection method is applied in the pre-filtering process to detect edges from the captured image, it will be less prone to get deceived by noise. Consequently, the system is able to analyze signs manipulated with rain, leaves and dirt and produce output that is quite accurate.

In the process of conducting the research work, we have identified a number of constraints and area of improvements. The most notable of them are listed as following.

- Limited Size of the training corpus
- Limitation of OCR for angled photos
- Image adjustment is not dynamic
- Overfitting of data from the Neural Network
- Machine translation needs optimization

Moreover, the authors would like to incorporate driver movement detection techniques through accelerometer, gyroscope and compass sensor data to align and compare that with the instruction from the traffic sign. Therefore, if the driver's movement data is conceived as illegitimate according to the traffic signs, the system will generate a warning sound.

Table 2: Demo Experimental Result.

Input Sentence Bangla	**Output Sentence English**
সামনে সরু সেতু আছে	Narrow Bridge
সামনে স্কুল	School
হাসপাতাল	Hospital
সামনে ওয়াই-জাংশন আছে	Y-Junction
থামানো নিষেধ	No Stopping
হর্ণ বাজানো নিষেধ	No Horn Honking
সর্বোচ্চ গতিসীমা	Highest Speedlimit
বিপদজনক খাঁদ	Dangerous Dip
বনভোজন এলাকা	Picnic Site

References

[1] Jack Greenhalgh, Majid Mirmehdi, "Recognizing Text-Based Traffic Signs," IEEE Transactions on Intelligent Transportation Systems, Volume 16 Issue 3, June 2015

[2] Swati M, K.V. Suresh, "Automatic traffic sign detection and recognition - A Review," 2017 International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET)

[3] J. Canny, "A Computational Approach to Edge Detection," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-8, no. 6, pp. 679-698, Nov. 1986.

[4] K. Ito, "Gaussian filter for nonlinear filtering problems," Proceedings of the 39th IEEE Conference on Decision and Control (Cat. No.00CH37187), Sydney, NSW, 2000, pp. 1218-1223 vol.2.

[5] Yan Zhang and Lee Makowski, "Auto-thresholding Edge Detector for bio-image processing," 2015 41st Annual Northeast Biomedical Engineering Conference (NEBEC), Troy, NY, 2015, pp. 1-2.

[6] A. Rehman, "Offline touched cursive script segmentation based on pixel intensity analysis: Character segmentation based on pixel intensity analysis," 2017 Twelfth International Conference on Digital Information Management (ICDIM), Fukuoka, 2017, pp. 324-327.

[7] Y. Li, Y. Fu, H. Li and S. Zhang, "The Improved Training Algorithm of Back Propagation Neural Network with Self-adaptive Learning Rate," 2009 International Conference on Computational Intelligence and Natural Computing, Wuhan, 2009, pp. 73-76.

[8] Linsen Yu, Yongmei Liu and Tianwen Zhang, "Using Example-Based Machine Translation Method For Automatic Image Annotation," 2006 6th World Congress on Intelligent Control and Automation, Dalian, 2006, pp. 9809-9812.

Proceedings, pages 61–65
Paris, UNESCO Headquarters, 5-6 December, 2019.

Automated Bangla Sign Language Conversion System: Present and Future

Professor Dr. Mohammad Nurul Huda, Ishteaque Alam, Sabbir Arif Siddique
eGenration Ltd.
mnh@cse.uiu.ac.bd, ishteaque.ark@gmail.com, sabb.a.sidd@gmail.com

Abstract

Deaf and dumb people communicate with each other through sign language. In this research work we have analyzed and acknowledged useful features from most prominent sign Language to Bangla text and speech language of classic and modern approaches.
Keywords: *Bangla Sign Language, Bidirectional System, Static,Dynamic, Gesture Recognition, Real time Detection*

Résumé

মূক ও বধির জনগোষ্ঠী পারস্পরিক যোগাযোগের জন্য ঈশারা ভাষা ব্যাবহার করে থাকে। আমাদের এই গবেষণাটি ঈশারা ভাষা কে বাংলা কথিত ও লেখ্য ভাষায় রুপান্তর করার বর্তমান ও ভবিষ্যৎ প্রযুক্তি নিয়ে আলোকপাত করেছে।

1. INTRODUCTION

The method of human communication, either spoken or written, consists of the use of words in a structured and conventional way. But the deaf and dumb people cannot hear or speak. So, deaf and dumb people around the world communicate using sign language as distinct from spoken language in their everyday lives. A Sign Language is a visual language that uses a system of manual, facial and body movements as the means of communication. Sign language is not a universal language, and different sign languages are used in different countries. The Broca's and Wernicke's areas of left hemisphere of the brain process sign languages like other natural languages [1].

Sign language, any means of communication through bodily movements, especially of the hands and arms, used when spoken communication is impossible or not desirable. Wherever vocal communication is impossible, as between speakers of mutually unintelligible languages or when one or more would be communicators are deaf, sign language can be used to bridge the gap.

There are about 70 million deaf [2] hearing-impaired as well as hearing people in the world who use sign language as their first language or mother tongue. Each country has one or sometimes two or more sign languages, although different sign languages can share the same linguistic roots in the same way as spoken languages do. Though it is not an international language, the universal features in sign languages help to make it possible for users of different sign languages to understand one another far more quickly than users of unrelated spoken languages can. Deaf children must have rights for equal and quality education like all other children and expect that their educational rights are respected and supported by educational authorities.

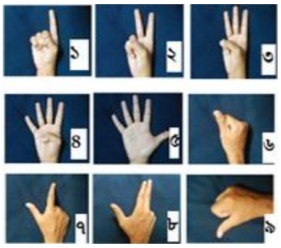


Fig. 1. Signs for Bangla numeral

Hundreds of sign languages are being used around the world, which include, American Sign Language (ASL), Japanese Sign Language, (JSL), British Sign Language (BSL), Austrian Sign Language (GS), Bangladeshi Sign Language (BdSL) and so on.

Figure-1 shows the Bangla numeral sign [3] and Figure-2 shows the two handed Bangla vowels sign [4].

Fig. 2. Two handed sign images for Bangla vowels

Nuances within the deaf communities are the reasons for most of the differences in these signs of that area, and has led to a significant evolution of sign language worldwide.

All over the world, many systems had been developed by the researchers for different sign languages. An interactive sign language teaching program [5] was developed by Kadam et al. for American Sign Language (ASL) using glove. For detecting the hand gesture, they used flex sensors in the fingers of the glove. Verma et al. developed a system [6] for detecting sign language gesture using Microsoft Kinect. Use of Microsoft Kinect provides highly strong features form depth image which are very useful to make an accurate gesture recognition system. A bidirectional translation system [7] for Japanese Sign Language (JSL) designed by Imagawa et al. They made a real-time system for tracking hands of signer and find the location of each hand which is very necessary for detecting sign language gestures. An application named 'Gesture Audio Video Conferencing Application' proposed by Tarte et al. in
[8]. This application model includes translation of Naturally Spoken English Sentences to Visual Sign Language Gesture and vice versa. In this project input hand gesture of sign language can be detected by gesture template matching and it will produce text to speech output. On the other hand, natural voice can be detected by voice recognition system to fetch the input sentence and producing output gestures for important words only.

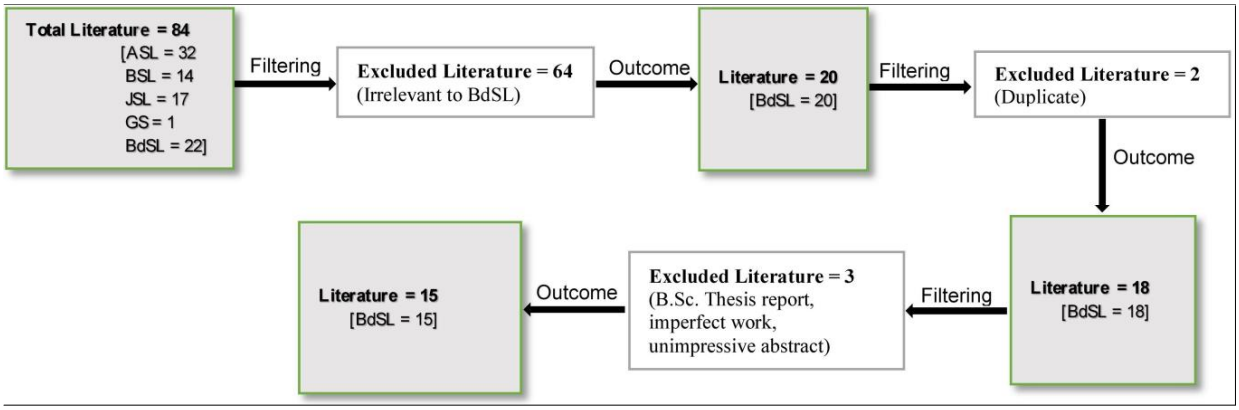


Fig. 3. Filtering articles to find all literature relevant to BdSL

Starner et al. developed two real-time hidden Markov model-based systems [9] for continuously detecting American Sign Language (ASL). Their first system used a camera on a desk which is actually second-person viewpoint for observing the signer's hand. The second system used a camera in a cap which was worn by the signer for observing his/her own hand gesture form first-person viewpoint. Using a 40 word lexicon, the first system achieved 92% accuracy while the second system achieved 98% accuracy.

The objective of the paper is to analysis the existing sign language translation system, a system that translate naturally Bangla spoken language to Bangla Sign Language and vice versa and then proposed a system that would have more flexibility and robustness for real life applications. For that purpose, we find out the previous related works which were developed for Bangla Sign Language in order to review and analyze these existing systems for developing a system to overcome the limitations of those.

Bangla is the most spoken language in Bangladesh and second in India. About 250 million native and about 300 million total speakers worldwide speaks in Bangla [10]. So a welldeveloped system for Bangla speech to Bangla Sign Language and vice versa is very much needed for communication with deaf people. While discussing further, we will use the term BL for naturally spoken Bangla Language and BdSL for Bangla Sign Language.

2. METHODOLOGY

To identify the best approach for a Sign Language translation system, one of the early steps is to review journal, articles, and conference papers related to this system. The best place for finding research papers are academic search engines, Google Scholar, Microsoft Academic Research, Directory of Open Access Journals, Science and Technology of Advanced Materials etc.

2.1 Search Methods

The results of search engines depend upon the selection of relevant keywords. In starting we used several different keywords, both alone and in combination. Later we narrow our focus and figured out which keywords describe our research topic best from the results of the searches. We also used available filters in the academic search engines to filter down to our relevant research papers. The results of academic search engines come in the form of an abstract with a full citation (author, journal title, volume, page numbers, year, etc.). We read title and abstract from the results and also keep noticed to the journal or conference title to select research papers of our research topic of this study. Most frequently used keywords for searching journal, articles and conference papers for BdSL were "Sign Language", "Recognition of Sign Language", "Hand Gesture Recognition", "Sign Language Interpreter", "Real-time Gesture Recognition", "Communication System with Deaf and Dumb People", "Interpreter for Deaf and Dumb People", "Bidirectional Communication System for Sign Language". Later for finding more specific research paper about BdSL, we used the keywords like "Bangla", "Bengali", "Bangladeshi Deaf and Dumb People" as a prefix before the most frequently used keywords mentioned above.

2.2 Selection Criteria

We used some selection criteria to find the best fit articles related to this system. Articles of only reputed journals and conferences relevant to the topic were considered. The filtering process is represented by the Figure-3.

We have selected only the articles which were published within the year of 2000 to 2015. Out of all these research papers we targeted only those which are related to Bangla Sign Language. After all the filtering process, finally we have 15 research works related to BdSL.

The selection and filtering process can be summarized as:

- The literature must have discussed about Sign Language.
- Since the topic is about BdSL, the literature has to contain discussion or proposal or implementation about Bangla Sign Language.
- PhD or MS research works have been prioritized.
- Relevant, unique and informatics works are included
- Duplicate literature are excluded.

3. RESULT OF RELATED RESEARCH PAPERS IN BANGLA

In total we have collected 22 research papers related to sign language to speech conversion and vice versa. Among them 15 research papers are related to Bangla Sign Language. In this study, we tried to discuss about the methodologies used for Bangla Sign language to Bangla spoken language conversion and vice versa. The communication system with deaf and dumb people using sign language can be discussed as one way communication and bi-directional communication.

3.1 One Way Communication

One way communication system can be either conversion of naturally spoken Bangla Language (BL) to Bangla Sign Language (BdSL) or conversion of BdSL to BL. All those approaches are discussed below.

3.1.1 *One Way Communication: Conversion of BL to BdSL*

: A human machine interface named ”Intelligent Assistant” designed by Eshaque et al. Which can understand only ten Bengali expressions. They used Microsoft Voice Command and Control Engine to capture sound input and convert it into recognizable specific text. Then the text is send to match with previously stored words in the system's knowledge base. If the word is found in the database then the system displays a pre-stored 3D graphical hand gesture according to that text. The accuracy of ”Intelligent Assistant” is 82% with only ten Bangla words. This system can understand only discrete words, not a whole sentence. They used one hand gesture but there are some signs in BdSL which requires both hands and facial expression too. Actually for understating only ten discrete words in Bangla, it doesn't require any intelligent system. Another system [12] developed by Sarkar et al. in 2009 which can translate an input Bangla sentence or word to the corresponding BdSL gestures using a set of rules. They took text input using either physical or on-screen keyboard. The translator dictionary is consist of 1000 words which were mostly taken from textbooks taught to children learning Bangla in the primary level. After taking an input and rearranging the words (if rearranging is necessary based on the rules), prestored video gestures of corresponding words are concatenated sequentially to produce a smooth output video in BdSL.

3.1.2 One Way Communication: Conversion of BdSL to BL

The conversion of BdSL to BL can be discussed based real time hand gesture recognition, one or two handed hand gesture recognition, static or dynamic gesture image recognition etc. Static gesture recognition means, a system can recognize sign gesture from a static image only. On the other hand dynamic gesture recognition means, the system can detect a sign gesture from real time video or any video input.

3.1.2.1 Real Time Dynamic BdSL Gesture Input

A geometrical model based hand gesture recognition system developed by Pavel et al. Which can analyze video clips of different BdSL gestures, processing them and gives an audio output. This system calculates angles between different parts of the hand with body and then matches those values with pre-stored values in database which they created manually for different sign expressions. Then audio file of the corresponding gesture is played as output. Another work under the same authors in 2004 can detect and simulate 9 different gestures. They have achieved 97% accuracy by testing 9 gestures with 20 different students. ”Intelligent Assistants for Speech Impaired People” developed by Rahman et al. which uses images of hand wearing glove as input. The glove contains different dots in each finger. The dots in the hand image are analyzed to detect which sign is shown by signer. Then pre-stored audio and written text output is shown corresponding to the sign. Only Bengali 1-10 numerals can be detected by this system. A real time computer vision based system made by Rahaman et al. that can detect two handed hand gesture showing sign from real time video recording and provides output for the corresponding BdSL sign. Only 6 Bengali vowels and 30 Bengali consonants can be recognized by this system. It was trained using 3600 images (36 different signs, 10 Signers-4 Female, 6 Male and 10 images for each). This system was tested using 3600 images of 10 different signers and achieved accuracy 98.17% for vowels and 94.75% for consonants. Another system developed by Jarman et al. in [3] can recognize 46 one hand gesture of Bengali alphabets and numerals. Multilayered feed-forward neural network with back-propagation algorithm was used to train the dataset. The system can perform rotation on BdSL gestures if it is required in some cases which is a good feature of this system. This system was tested using dynamic sign as 3 seconds videos where frames were extracted at a rate of 15 fps. The average recognition rate this system is 88.69%. As the system can perform sign detection from video clips which we called dynamic sign detection, it can easily be convert to perform in real time sign detection.

3.1.2.2 Recognition of Static Two Handed Hand Gesture

An intelligent sign language verification system offered by Rahim et al. used image processing, clustering and neural network concepts. The database of this system contains clustering information for the detection of gestures. Useful features are generated from input image by image processing technique. These features are classified by Neural Networking process according to the clusters information. Then the output is shown both in visual and textual form. This system was tested on BdSL but it can any sign language with prior image processing and clustering. Not only two handed hand gesture, it is even capable of deal with other body parts according to the training. This system is called ”Intelligent System” because it can train itself automatically. They achieved 92% (approx.) accuracy by using 24 different sign images with 10 clusters. Using artificial neural network, Rahman et al. designed a system described in. This system does not require any gloves or visual marking system, it just requires image of bare hand. It can recognize 36 letters of BdSL alphabet with an accuracy of 80.902%. Deb et al. developed a system [4] which used two different color wristbands to easily remove the forearm in binary image. After detecting the hand sign from the candidate region, this system used template matching with the use of normalizes cross-correlation to recognize different types of hand sign. This

system can detect Bangla Sign Language gestures of 10 Bangla alphabets with 96% success rate. For detecting two handed hand gesture, Yasir et al. designed a system which can recognize Bangla sign images of 15 Bangla letters. They used machine learning techniques PCA and LDA, neural network for training and testing purpose of their system. To find out a better solution they used three different models. In first model, input images were trained and tested using neural network only and resulted output comes with an accuracy of 67.15%. In the second model, Principal component analysis was used to train the dataset and resulted output had an accuracy of only 26%. And in last model LDA algorithm was used for feature extraction and resulted output achieved an accuracy of 100%.

3.1.2.3 Recognition of Static Single Hand Gesture:

A fuzzy rule based system designed by Ayshee et al. to recognize static single hand gesture for only two Bengali letters. Angles of fingers are calculated to determine the hand configurations which are basically the fuzzy rules to match them with predefined fuzzy rules for two Bengali letters. In this system, static image of single hand gesture can be taken as input from one viewpoint only. All the fingers may not be visible from one viewpoint which is a limitation of this system. Karmokar et al. developed a system in by using Neural Network Ensemble for the recognition of single hand gesture of BdSL. This system was trained for 47 signs of BdSL including Bengali alphabets and numerals. They achieved 93% accuracy using negative correlation learning with 10 neural networks with feature extraction.

3.1.2.4 Recognition of Static Image of BdSL:

For recognition of static images, a system in developed by Rahim et al. which can take input as static sign gesture image and find out grammatical similarity with previously learned images. This system used back propagation algorithm of artificial neural network for learning and detection purpose. With 2000 learned images, the system was tested and the output achieved 99% accuracy.

3.2 Bi-directional Communication:

A bidirectional system proposed by Datta et al. based on their previous work. This two way communication system includes Bangla Text to BdSL and BdSL to Bangla Text conversion. They did not do any implementation work for this proposed two way communication system but their previous work has an implementation for the conversion of Bangla text to BdSL gestures.

4. PROPOSED SYSTEM

For fluent communication with deaf people who use BdSL, our proposed system contains two major parts. This system includes conversion of (a) Bangla Language to Bangla Sign Language and (b) Bangla Sign Language to Bangla Language.

4.1 Bangla Language to Bangla Sign Language

The basic steps for the conversion of Bangla Language to Bangla Sign Language can be divided into four parts. Those are taking input as text, processing input based on predefined grammatical rules, searching processed text into database, preparing output and display it in monitor. The whole process can be simply represented by the Figure-4. First of all we have to construct a database containing all Bangla alphabets, numerals, words used in BdSL and mapping them with their corresponding BdSL gestures. BdSL gestures can be made as a mini video clips for each sign.

Sentence making grammar for Bangla Language and Bangla Sign Language has a huge difference. Both of them follow their own grammatical rules for making a complete understandable sentence. So for the conversion of Bangla Language to BdSL and vice versa, we have to define some rules for this type of conversion and implement it properly. For learning rules and implement it, we can use machine learning process. Rules based on Bangla Parts of Speech can also be implemented.

The input of Bangla Language can be either voice input using microphone or text input using physical or onscreen keyboard. Still there is no well-developed Bangla voice recognition system available and to stay focused in our main purpose, we will simply use textual input of Bangla language.

The output will consist of two parts. One part will show a textual output of grammatically converted BdSL from Bangla Language if the conversion is needed. Other part will show the corresponding BdSL gestures output of that converted BdSL text. The BdSL gestures will be represented by sequentially joining pre-stored video clips of corresponding words in that BdSL text.

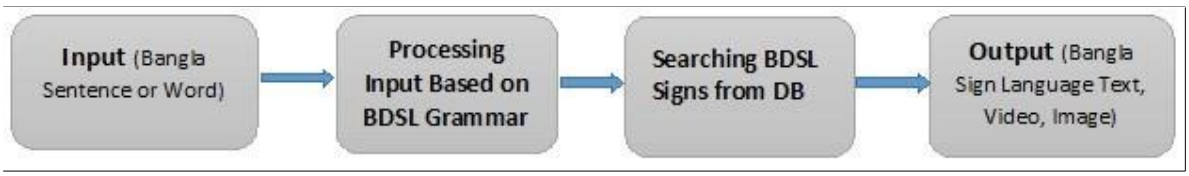


Fig. 4. Conversion of Bangla Spoken Language to BdSL

So the overall process is, we will take Bangla Language as textual input. The input text will rearranged by the grammatical rules of Bangla Language to Bangla Sign Language conversion if the conversion is needed to perform. The words in the converted BdSL text will be searched in the database to extract their corresponding gesture video clips. After sequentially joining the video clips for each words in that text, both the textual and video output for BL to BdSL will be shown in the output screen. A similar process has been developed in [12] which has a translator dictionary containing 1000 words. A complete set of words used in BdSL need to add with their

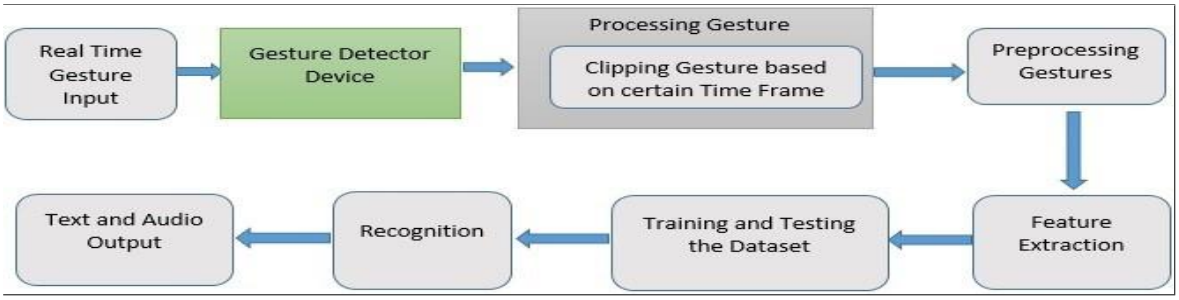


Fig. 5. Conversion of BdSL to Bangla Spoken Language

corresponding BdSL gestures. They implemented few rules for the conversion of BL to BdSL where many more rules need to be added.

4.2 Bangla Sign Language to Bangla Language

The target of BdSL to Bangla Language conversion in our proposed system is to detecting real time continuous dynamic signs using gesture detecting devices like leap motion or Kinect sensor and producing output in Bangla Language. Major parts of this conversion are taking real time gesture input, processing the gesture to extract features, training and testing input data, finally producing the recognized text and audio for the input Bangla gesture sign. Conversion of Bangla Sign Language to Bangla Language can be represented by the
Figure-5.

The main purpose of this proposed model is to detecting consecutive Bangla sings in real time. The input real time gesture can be clipped based on a fixed time frame. If we use certain time frame to clip the gesture input then we have to search for a single unique sign in that sign gesture. The gesture sign will pass through the preprocessing step. Only hand signs are considering for the detection of BdSL, so we have to detect the hand region from extracted gestures, crop that hand region. Then important features have to extract from that gesture sign. Initially we have to train our dataset using those features to detect various BdSL gestures. After the training session, an input hand gesture sign can be recognized by comparing its features with the trained dataset. The output will be shown as both in textual and audio format for the corresponding input Bangla gesture sign.

5. CONCLUSION

This paper presents a comparison among the research works had done for development of the communication system between Bangla Sign Language and Bangla Language. It also proposed a two way communication system for BdSL to Bangla Language and vice versa. We focused on real time consecutive dynamic Bangla sign recognition for BdSL to Bangla Language conversion. On the other hand, Bangla Language to BdSL conversion needs a huge dataset of Bangla words mapped with their corresponding Bangla signs. A large set of grammatical rules has to develop and implement for proper conversion of Bangla Language to BdSL. Such a welldeveloped two way communication system is not still available for the deaf people who use BdSL in their regular life. In this system, we tried to pick the best solutions for the limitations identified in the previous works and combine them with our own proposal.

Reference

[1] R. Campbell, M. MacSweeney, and D. Waters, “Sign language and the brain: a review,” Journal of Deaf Studies and Deaf Education, 2007.

[2] “Sign language - wfd — world federation of the deaf,” http://wfdeaf. org/human-rights/crpd/sign-language, (Visited on 09/30/2015).

[3] N. A. M. J. I. Angur M. Jarman, Samiul Arshad, “An automated bengali sign language recognition system based on fingertip finder algorithm,” in International Journal of Electronics & Informatics (IJEI), ser. 1, vol. 4. Center for Natural Science & Engineering Research, 2015.

[4] K. Deb, H. P. Mony, and S. Chowdhury, “Two-handed sign language recognition for bangla character using normalized cross correlation,” Global Journals Inc.(USA), vol. 12, 2012.

[5] K. Kadam, R. Ganu, A. Bhosekar, and S. Joshi, “American sign language interpreter,” in Technology for Education (T4E), 2012 IEEE Fourth International Conference on. IEEE, 2012, pp. 157–159.

[6] H. V. Verma, E. Aggarwal, and S. Chandra, “Gesture recognition using kinect for sign language translation,” in Image Information Processing (ICIIP), 2013 IEEE Second International Conference on. IEEE, 2013, pp. 96–100.

[7] K. Imagawa, S. Lu, and S. Igi, “Color-based hands tracking system for sign language recognition,” in Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on. IEEE, 1998, pp. 462–467.

[8] N. T. Tarte, S. N. Bhadane, and P. S. Kulkarni, “A gesture audio-video conferencing application for the ease of communication between normal person and deaf and dumb person,” Int. J. Sci. Eng. Technol. Res., vol. 3, no. 5, p. 14871490, 2014.

[9] T. Starner, J. Weaver, and A. Pentland, “Real-time american sign language recognition using desk and wearable computer based video,” Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 20, no. 12, pp. 1371–1375, 1998.

[10] “Bengali language - wikipedia, the free encyclopedia,” https://en. wikipedia.org/wiki/Bengali language, (Visited on 09/30/2015).

Proceedings, pages 66–70
Paris, UNESCO Headquarters, 5-6 December, 2019.

Bangla Phonetic Features Extraction for Automatic Speech Recognition

Professor Dr. Mohammad Nurul Huda[1], Sabbir Arif Siddique[2], Ishteaque Alam[3]
Department of Computer Science and Engineering
United International University and eGeneration Ltd.
Dhaka-1212, Bangladesh
mnh@cse.uiu.ac.bd[1], sabb.a.sidd@gmail.com[2], ishteaque.ark@gmail.com[3]

Abstract

This research constructs a distinctive phonetic feature (DPF) table for all the phonemes pronounced in Bangla (widely known as Bengali) language where the whole study is divided into two parts. In the first part, a DPF table is constructed, while the second part deals with Bangla automatic speech recognition (ASR) using DPFs. For Bangla language, fifty three phonemes including both vowels and consonants are considered in which the phones, শ (/s/) and স (/s/), and, ণ (/n/) and ন (/n/) contain approximately same spectrum and hence, they share same DPFs. In the DPF table, twenty two DPFs (Silence, Short Silence, Stop, …) are required for representing all the Bangla phonemes. On the other hand, the second part comprised of three stages: i) first stage deals with acoustic features, mel frequency cepstral coefficients (MFCCs) extraction, ii) second stage embeds DPFs extraction procedure using a multilayer neural network (MLN) and iii) the final stage integrates a triphone-based hidden Markov model (HMM) for generating the output text strings by inputting log values of twenty two dimensional DPFs. In the experiments on Bangla Newspaper Article Sentences, it is observed that the DPF-based ASR system provides higher word correct rate, word accuracy and sentence correct rate in comparison with the standard MFCC-based method.

Keywords: Distinctive phonetic feature; mel frequency cepstral coefficient; multilayer neural network; automatic speech recognition; hidden Markov model

Résumé

এই গবেষণাটি বাংলা (ব্যাপকভাবে বাংলা হিসাবে পরিচিত) ভাষায় উচ্চারণ করা সমস্ত ফোনের জন্য একটি স্বতন্ত্র ফোনেটিক বৈশিষ্ট্য (ডিপিএফ) সারণী তৈরি করে যেখানে পুরো অধ্যয়ন দুটি অংশে বিভক্ত। প্রথম অংশে একটি ডিপিএফ টেবিল তৈরি করা হয়েছে, দ্বিতীয় অংশে ডিপিএফ ব্যবহার করে বাংলা স্বয়ংক্রিয় কথা থেকে লেখা (এএসআর) নিয়ে আলোচনা করা হয়েছে। বাংলা ভাষার জন্য, স্বর এবং ব্যঞ্জন উভয় সহ পঞ্চাশটি ফোনেম বিবেচনা করা হয় যার মধ্যে ফোনগুলি, শ (/s/) এবং স (/s/), এবং, " ণ (/n/) এবং" ন (/ n /) প্রায়ই থাকে এবং একই স্পেকট্রাম ব্যবহার করে এবং তাই তারা একই ডিপিএফ ভাগ করে নেয়। ডিপিএফ টেবিলে, সমস্ত বাংলা ফোনের প্রতিনিধিত্ব করার জন্য বাইশটি ডিপিএফ (সাইলেন্স, শর্ট সাইলেন্স, স্টপ,…) প্রয়োজন। অন্যদিকে, দ্বিতীয় অংশটি তিনটি পর্যায়ের সমন্বয়ে গঠিত: i) প্রথম পর্যায়ে অ্যাকোস্টিক বৈশিষ্ট্যগুলি নিয়ে কাজ করে, মেল ফ্রিকোয়েন্সি সিপস্ট্রাল সহগ (এমএফসিসি) বের করে, ii) দ্বিতীয় পর্যায়ে মাল্টিলেয়ার নিউরাল নেটওয়ার্ক (এমএলএন) ব্যবহার করে এবং iii) চূড়ান্ত পর্যায়ে ডিপিএফগুলি নিষ্কাশন প্রক্রিয়া এম্বেড করা হয় বাইশ মাত্রিক ডিপিএফ-এর লগ মানগুলি ইনপুট করে, আউটপুট স্ট্রিং উৎপন্ন করার জন্য একটি ট্রাইফোন ভিত্তিক হিডেন মার্কোভ মডেল (এইচএমএম) ব্যবহার করে। বাংলা সংবাদপত্রের নিবন্ধের বাক্যগুলির পরীক্ষায় দেখা গেছে যে ডিপিএফ-ভিত্তিক এএসআর সিস্টেম স্ট্যান্ডার্ড এমএফসিসি ভিত্তিক পদ্ধতির তুলনায় উচ্চতর শব্দ সঠিক হার, শব্দের যথার্থতা এবং বাক্য সঠিক হার সরবরাহ করে।

1. Introduction

There have been many literatures in automatic speech recognition (ASR) systems for almost all the major languages in the world. Unfortunately, only a very few works have been done in ASR for Bangla (can also be termed as Bengali), which is one of the largely spoken languages in the world. More than 220 million people speak in Bangla as their native language. It is ranked seventh based on the number of speakers [1]. A major difficulty to research in Bangla ASR is the lack of proper speech corpus. Some efforts are made to develop Bangla speech corpus to build a Bangla text to speech system [2]. However, this effort is a part of developing speech databases for Indian Languages, where Bangla is one of the parts and it is spoken in the eastern area of India (West Bengal and Kolkata as its capital). But most of the natives of Bangla (more than two thirds) reside in Bangladesh, where it is the official language. Although the written characters of standard Bangla in both the countries are same, there are some sounds that are produced variably in different pronunciation of Standard Bangla, in addition to the the myriad of phonological variations in non-standard dialects [3]. Therefore, there is a need to do research on the main stream of Bangla, which is spoken in Bangladesh, ASR.

Some developments on Bangla speech processing or Bangla ASR can be found in [4]-[11], where various hidden Markov model (HMM)-based ASR systems have been developed. Most of these ASR systems make use of a preprocessed form, such as mel-frequency cepstral coefficients (MFCCs), of the speech signal, which encodes the time-frequency distribution of signal energy. However, these MFCC-based systems do not provide better recognition performance in real acoustic conditions (See Figure 1(a)). On the other hand, a system based on Distinctive Phonetic Features (DPFs) exhibits higher recognition accuracy in practical conditions and models coarticulatory phenomena more naturally [12](See Figure 1(b)). From the Figures 1(a) and 1(b), it is shown that the DPF-based system outputs few misclassifications. The main problem for the Bangla language is that DPF table is yet to be constructed.

In this paper, we have designed a Distinctive Phonetic Feature (DPF) table for all the phonemes pronounced in Bangla language. The first part of the research deals with a DPF table construction, while the second part constructs a Bangla ASR using DPFs. In the DPF table, twenty two

DPFs are required for representing all the Bangla phonemes. On the other hand, the second part comprised of three stages: i) first stage deals with acoustic features, mel frequency cepstral coefficients (MFCCs), extraction, ii) second stage embeds DPFs extraction procedure using a multilayer neural network (MLN) and iii) the final stage integrates a triphone-based HMM for generating the output text strings by inputting log values of twenty two dimensional DPFs.

The paper is organized as follows. Section II briefly describes an approximate phonetic scheme and speech corpus for Bangla and formation of words, and speech corpus for Bangla. Section III explains about Bangla DPFs, while Section IV deals with Proposed ASR construction using Bangla DPFs. Again, Section V gives experimental setup, results and discussion on Bangla continuous word recognition. Finally, Section VI draws some conclusions with future directions.

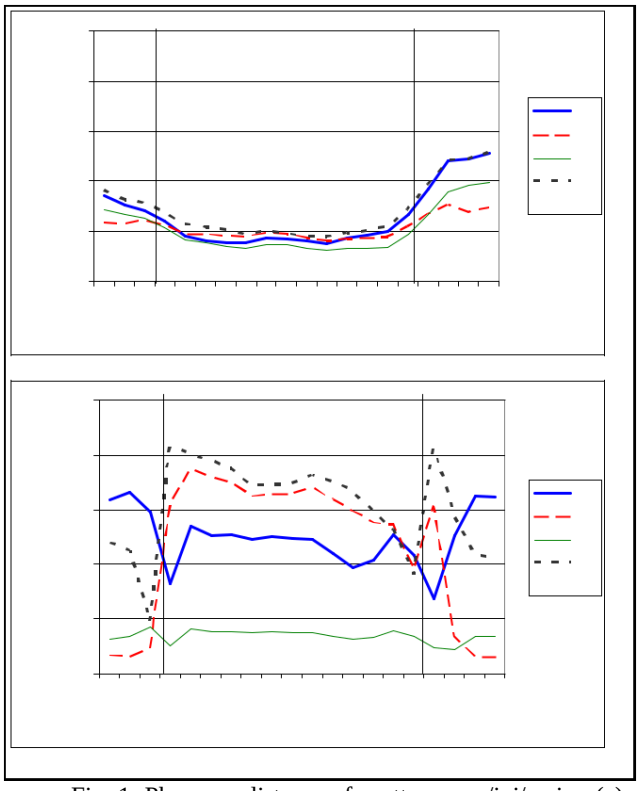


Fig. 1. Phoneme distances for utterance, /ioi/ using (a) MFCC-based system and (b) DPF-based system.

2. PHONETIC SCHEME AND CORPUS FOR BANGLA

2.1 Bangla Phonemes

The unprotected PDF files will appear in the on-line proceedings directly as received. Do not print the page Citing References in the Text. The phonetic inventory of Bangla consists of 14 vowels, including seven nasalized vowels, and 29 consonants. An approximate phonetic scheme in IPA is given in [13][14], where only the main 7 vowel sounds are shown, though there exists two more long counterpart of /i/ and /u/, denoted as /i:/ and /u:/, respectively. These two long vowels are seldom pronounced differently than their short counterparts in modern Bangla. There is controversy on the number of Bangla consonants.

2.2 Bangla Words

TABLE I. EXAMPLES OF SOME BANGLA WORDS WITH THEIR IPA

Bangla Word	English Pronunciation	IPA	Our Symbol
আমরা	AAMRA	/a m r a/	/aa m r ax/
আচরণ	AACHORON	/a tʃ r n/	/aa ch ow r aa n/
আবেদন	ABEDON	/a b æ d̪ n/	/ax b ae d aa n/

Table I lists some Bangla words with their written forms and the corresponding IPA. From the table, it is shown that the same 'আ' (/a/) has different pronunciation based on succeeding phonemes 'ম' /m/, 'চ' /tʃ/ and 'ব' /b/. Thesepronunciations are sometimes long or short. For long and short 'আ' we have used two different phonemes /aa/ and /ax/, respectively. Similarly, we have considered all variations of same phonemes and consequently, found total 51 phonemes excluding beginning and end silence (/sil/) and short pause (/sp/).

2.3 Bangla Speech Corpus

Hundred sentences from the Bengali newspaper "Prothom Alo" [15] are uttered by 30 male speakers of different regions of Bangladesh. These sentences (30x100) are used as training corpus (D1). On the other hand, different 100 sentences from the same newspaper uttered by 10 different male speakers are used as test corpus (D2). All of the speakers are Bangladeshi nationals and native speakers of Bangla. The age of the speakers ranges from 20 to 40 years. We have chosen the speakers from a wide area of Bangladesh: Dhaka (central region), Comilla – Noakhali (East region), Rajshahi (West region), Dinajpur – Rangpur (North-West region), Khulna (South-West region), Mymensingh and Sylhet (North-East region). Though all of them speak in standard Bangla, they are not free from their regional accent.

3. PROPOSED BANGLA PHONETIC FEATURES

A phoneme can easily be identified by its DPFs [16][17]. In this paper we have proposed Bangla DPFs for all the phonemes with their international phonetic alphabet (IPA) and Bangla orthographic transcription. Here, the fifty three Bangla phonemes and twenty two DPFs for each phoneme are silence, short silence, stop, nasal, bilabial, fricative, liquid, lenis, vowel, front, central, back, unvoiced, long, short, dipthong, high, low, medium, round, unround and glottal, which shown in the table horizontally and vertically, respectively. Here, (Front, Back, Central) and (High, Low, Medium) represent tongue position in forward and backward, and upward and downward directions, respectively. Besides, plus (+) and minus (-) elements in the table represent whether corresponding element is present or absent, respectively.

4. PROPOSED ASR SYSTEM USING DPFS

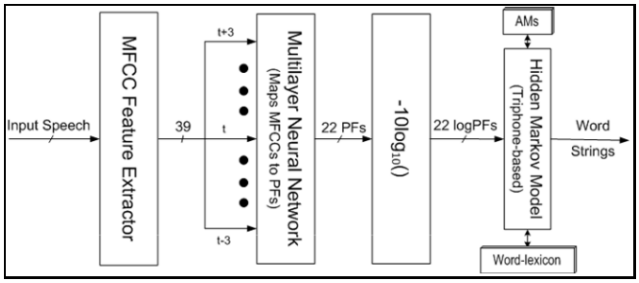


Fig. 2. Proposed PF-based ASR System.

We have implemented a DPF-based ASR system with an input acoustic vector of MFCCs using an MLN which is shown in Figure 2. This system comprised of three stages: i) first stage deals with acoustic features, MFCCs extraction, ii) second stage embeds DPFs extraction procedure using an MLN and iii) the final stage integrates a triphone-based HMM for generating the output text strings by inputting logarithmic values [17] of twenty two dimensional DPFs. The thirty nine dimensional MFCCs extracted in the first stage are entered into the MLN with five layers including three hidden layers after combining a current frame xt with the other two frames that are three points before and after the current frame (xt-3, xt+3) where the MLN generates twenty two DPF values for each input frame of 39x3 features. The three hidden layers comprised of 400, 200 and 100 units, respectively. The MLN is trained using the standard back-propagation algorithm.

5. Experiments

5.1 Setup

For evaluating word recognition performance, word correct rate (WCR), word accuracy (WA) and sentence correct rate (SCR) for D2 data set are evaluated using an HMM-based classifier. The D1 data set is used to design Bangla triphone HMMs with five states, three loops, and left-to-right models. Input features for the classifier are 39 dimensional MFCCs and log values of 22 dimensional PFs. The mixture components are set to 1, 2, 4 and 8.

For evaluating the performance of standard MFCC-based method including the proposed method, we have designed the following experiments:

(a) MFCC:dim-39 [Baseline]

(b) PF:dim-22 [Proposed]

In our experiments the range of output is from 0 to 1, where the non-linear function is a sigmoid, (1/(1+exp(-x))) for the hidden and output layers of MLN. For evaluating PF correct rate we have considered 0.20 as threshold to obtain better segmentation. Here, 0.20 is considered as threshold by observing the experimental results.

5.2 Result Analysis and Discussion

Segmentation for silence, short silence, stop, nasal, bilabial, fricative, liquid, lenis, vowel, front, central, back, unvoiced, long, short, dipthong, high, low, medium, round, unround and glottal PFs are depicted in Figures 3 and 4 for ideal and real cases for utterance, /prothom/. From both the figures, it is observed that segments of nasal, liquid, vowel and front are more precise (follows ideal line) in Figure 3, and unvoiced, long, dipthong, high, low, medium, unround and glottal exhibit better segments with respect to ideal segmentation in Figure 4. Again, Figure 5 shows correct rates for each of the DPFs using the test utterances in D2 data set, where DPF correct rates for the corresponding DPFs are 97.83%, 52.88%, 75.15%, 75.88%, 64.30%, 84.68%, 49.20%, 84.67%%, 95.72%, 87.83%, 88.22%, 78.42%, 93.79%, 87.49%, 86.65%, 82.97%, 77.82%, 70.75%, 92.62%, 86.15%, 89.00%, and 100.00%, respectively.

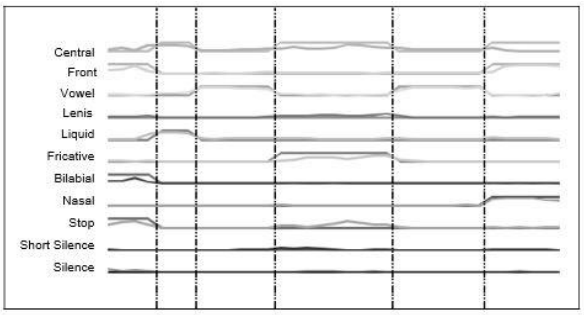


Fig. 3. Segmentation for silence, short silence, stop, nasal, bilabial, fricative, liquid, lenis, vowel, front and central DPFs using the utterance /prothom/.

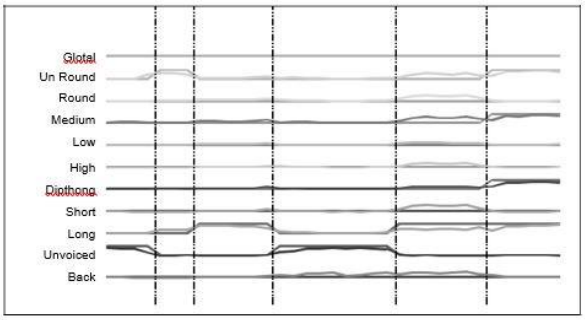


Fig. 4. Segmentation for back, unvoiced, long, short, dipthong, high, low, medium, round, unround and glottal DPFs using the utterance /prothom/.

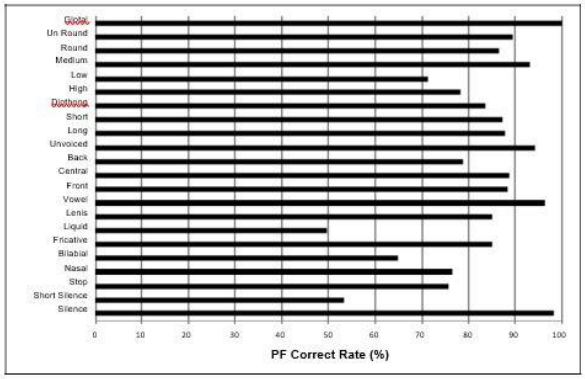


Fig. 5. Correct rates (%) for silence, short silence, stop, nasal, bilabial, fricative, liquid, lenis, vowel, front, central,

back, unvoiced, long, short, dipthong, high, low, medium, round, unround and glottal DPFs using the test utterances in D2 data set.

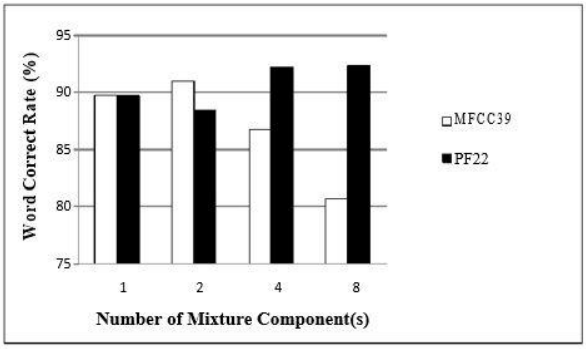


Fig. 6. Word Correct Rates for MFCCs and Proposed Method.

Figure 6 shows the comparison of word correct rates among all the investigated methods, standard MFCC-based method and proposed method. Among all the mixture components except two, the proposed method shows higher correctness in comparison with baseline. It is observed from the figure that the proposed method exhibits its best performance (92.25%) at mixture component eight. Besides, the mixture components, four and eight in the proposed method exhibit almost the same performance. Therefore, further investigation for higher correctness in higher mixture component is not required.

Word accuracies for the different investigated mixture components in standard MFCC-based and proposed methods are depicted in Figure 7. In mixture components one, two, four and eight, the proposed method provides 89.45%, 88.02%, 91.43% and 91.64% accuracies respectively, whereas 89.03%, 90.33%, 86.17% and 80.43% are observed in baseline method for the corresponding mixture components respectively.

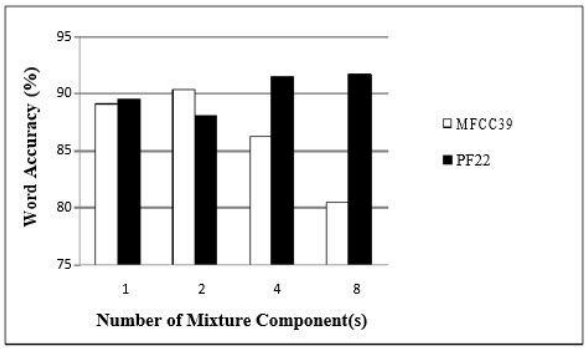


Fig. 7. Word Accuracies for MFCCs and Proposed Method.

Sentence correct rate which is shown in Figure 8 gives an idea about the performance of ASR systems investigated. For the experimented mixture components, there are 89.20%, 88.20%, 91.50% and 91.60% SCRs are found in the proposed method respectively, while baseline system generates 88.60%, 90.00%, 85.00% and 79.20% for the same experimental conditions.

Table II exhibits word recognition performance with respect to correctly recognized words (H), deletion (D), substitution (S) and insertion (I), respectively for the

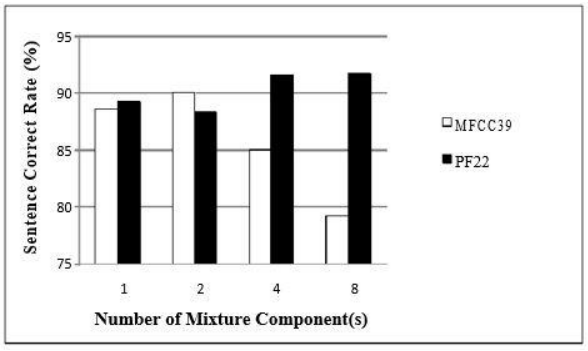


Fig. 8. Sentence Correct Rates for MFCCs and Proposed Method.

experimented mixture components in both the investigated ASR systems using the 3290 input words. For proposed and baseline methods, the H, D, S and I are 3035, 52, 203 and 20 respectively; and 2654, 202, 434 and 8, respectively for the investigated mixture component eight. Here, proposed method inserted more words than baseline. On the other hand, sentence recognition information for the investigated mixture components is provided in Table III using 1000 input spoken sentences.

TABLE II. WORD INFORMATION FOR INVESTIGATED METHODS WHERE H, D, S AND I REPRESENT CORRECT WORDS, DELETION, SUBSTITUTION AND INSERTION OUT OF 3290 RESPECTIVELY

		Mixture Components			
		Mix 1	Mix 2	Mix 4	Mix 8
MFCC 39	H	2950	2992	2851	2654
	D	91	75	114	202
	S	249	223	325	434
	I	21	20	16	8
AF 22	H	2952	2908	3033	3035
	D	77	106	58	52
	S	261	276	199	203
	I	9	12	25	20

TABLE III. SENTENCE INFORMATION FOR INVESTIGATED METHODS WHERE H, AND S REPRESENTS CORRECTLY AND INCORRECTLY RECOGNIZED SENTENCES RESPECTIVELY OUT OF 100

		Mixture Components			
		Mix 1	Mix 2	Mix 4	Mix 8
MFCC 39	H	886	900	850	792
	S	1144	100	150	208
PF 22	H	892	882	915	916
	S	108	118	85	84

6. Conclusion

This paper has constructed a distinctive phonetic feature table for Bangla automatic speech recognition. In the first part of the research twenty two phonetic features are considered for Bangla spoken language and the second part of the research designs an ASR system using the DPFs considered here. The following conclusions are given:

(i) Segmentation for each of the DPFs follows ideal boundaries for an input spoken sentence.

(ii) Correct rates for most of the DPFs are above 80%.

(iii) Word correct rate, word accuracy and sentence correct rate for the proposed method using all the investigated mixture components except two are better in comparison with the standard MFCC-based method.

In near future, the author would like to evaluate DPFs using recurrent neural network (RNN), which accommodates longer context window in its architecture. Besides, Deep Learning will be integrated for Bangla Speech Recognition. Moreover, the authors evaluate the experiments for gender independent environments.

7. Acknowledgements

This work was powered by United International University and eGeneration Ltd. jointly.

8. References

[1]http://en.wikipedia.org/wiki/List_of_languages_by_total_speakers, Last accessed April 11, 2009.

[2]S. P. Kishore, A. W. Black, R. Kumar, and Rajeev Sangal, "Experiments with unit selection speech databases for Indian languages," Carnegie Mellon University.

[3]http://en.wikipedia.org/wiki/Bengali_phonology, Last accessed April 11, 2009.

[4]S. A. Hossain, M. L. Rahman, and F. Ahmed, "Bangla vowel characterization based on analysis by synthesis," Proc. WASET, vol. 20, pp. 327-330, April 2007.

[5]M. A. Hasnat, J. Mowla, and Mumit Khan, " Isolated and Continuous Bangla Speech Recognition: Implementation Performance and application perspective, " in Proc. International Symposium on Natural Language Processing (SNLP), Hanoi, Vietnam, December 2007.

[6]R. Karim, M. S. Rahman, and M. Z Iqbal, "Recognition of spoken letters in Bangla," in Proc. 5th International Conference on Computer and Information Technology (ICCIT02), Dhaka, Bangladesh, 2002.

[7]A. K. M. M. Houque, "Bengali segmented speech recognition system," Undergraduate thesis, BRAC University, Bangladesh, May 2006.

[8]K. Roy, D. Das, and M. G. Ali, "Development of the speech recognition system using artificial neural network," in Proc. 5th International Conference on Computer and Information Technology (ICCIT02), Dhaka, Bangladesh, 2002.

[9]M. R. Hassan, B. Nath, and M. A. Bhuiyan, "Bengali phoneme recognition: a new approach," in Proc. 6th International Conference on Computer and Information Technology (ICCIT03), Dhaka, Bangladesh, 2003.

[10]K. J. Rahman, M. A. Hossain, D. Das, T. Islam, and M. G. Ali, "Continuous bangle speech recognition system," in Proc. 6th International Conference on Computer and Information Technology (ICCIT03), Dhaka, Bangladesh, 2003.

[11]S. A. Hossain, M. L. Rahman, F. Ahmed, and M. Dewan, "Bangla speech synthesis, analysis, and recognition: an overview," in Proc. NCCPB, Dhaka, 2004.

[12]K. Kirchhoff, et. al, "Combining acoustic and articulatory feature information for robust speech recognition," Speech Commun.,vol.37, pp.303-319, 2002.

[13]C. Masica, The Indo-Aryan Languages, Cambridge University Press, 1991.

[14]Ghulam Muhammad, Yousef A. Alotaibi and Mohammad Nurul Huda, "Automatic Speech Recognition for Bangla Digits," ICCIT'09, Dhaka, Bangladesh, December 2009.

[15]Daily Prothom Alo. Online: www.prothom-alo.com

[16]S. King and P. Taylor, "Detection of Phonological Features in Continuous Speech using Neural Networks," Computer Speech and Language 14(4), pp. 333-345, 2000.

[17]S. King, et. al, "Speech recognition via phonetically features syllables," Proc ICSLP'98, Sydney, Australia, 1998.

[18]T. Fukuda and T. Nitta, "Noise-robust ASR by Using Distinctive Phonetic Features Approximated with Logarithmic Normal Distribution of HMM," Proc. Eurospeech 2003, Vol.III, pp.2185-2188, Sep. 2003.

Proceedings, pages 71–74
Paris, UNESCO Headquarters, 5-6 December, 2019.

Russian Sign Language: History, Grammar and Sociolinguistic Situation in Brief

Ildar Kagirov, Dmitry Ryumin, Denis Ivanko, Alexander Axyonov, Alexey Karpov
St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences
SPIIRAS, 39, 14th Line, St. Petersburg, 199178, Russian Federation
kagirov@iias.spb.su, karpov@iias.spb.su

Abstract

The paper briefly sketches some outlines of Russian sign language: vocabulary, grammar, dialects and speech styles. Based on a range of studies on Russian sign languages published in 21st century, it aims to provide the reader with a representative list of recent papers, monographs and electronic resources about Russian sign language and a general notion of it. Besides, this paper makes some contribution to the description and documentation of Russian sign language, because nowadays, the Russian sign language remains a low-resource language just for spoken communication of deaf people in Russia and neighbouring countries, and it is still poorly studied.

Keywords: Russian sign language, sign language grammar, signed Russian

Résumé

В данной статье конспективно излагаются базовые сведения о русском жестовом языке: словари, грамматика, диалекты и речевые регистры. Работа основана на ряде исследований, опубликованных с начала 21 века, и призвана снабдить читателя перечнем основных материалов по русскому жестовому языку, а также дать общее представление о нем. Кроме того, статья вносит некоторый вклад в описание и документацию русского жестового языка, поскольку на сегодня русский жестовый язык все еще остается малоресурсным языком устного общения глухих людей в России и близлежащих стран, и он пока недостаточно изучен.

1. Introduction

This paper surveys basic features of Russian sign language (RSL), focusing on genetic classification, history, dialectal variations and elements of phonology and grammar. The aim is to give the reader a notion of RSL, as well as provide him with a list of essential papers and monographs about RSL.

RSL is the language of communication among the deaf and hard of hearing in Russia and some neighbor countries (mainly ex-Soviet countries; the main exception is Bulgaria). The total number of people using RSL in everyday life is more than 120 thousand (according to Ethnologue). Since 2012, RSL has enjoyed an official status in Russia as a language of communication.

2. Brief History and Genetic Classification of Russian Sign Language

The first school for the deaf in Russia was founded in 1806 near St. Petersburg (Williams and Fyodorova, 1993). The first deaf teachers came from Europe, and schools in St. Petersburg and Moscow maintained close ties with other European deaf schools until 1917. It is traditionally believed that RSL belongs to the French sign language family. However, this statement is based only on the fact that the first sign language teachers in Russia - Jean-Baptiste Jauffret and father Sigmund - came from France and Austria, respectively. Little is known about their teaching methods (Williams and Fyodorova, 1993).

This point of view is followed by the author of the study (Bickford, 2005), who provides analysis of wordlists in various sign languages, which does not confirm the hypothesis about the relationship between Russian and French sign languages.

3. Dialects and Speech Registers in RSL

There are not many studies on RSL dialects. Almost all the researchers note that there are certain differences of the phonological and lexical nature between idioms of RSL. However, in the study (Burkova and Varinova, 2012), it was shown that the real level of lexical differences in the local varieties of RSL is much lower than is commonly believed. Researches on grammatical differences in the dialects of RSL are very scarce, with very few exceptions. One can mention the work (Kimmelman, 2009), in which, inter alia, problems of reflexive pronouns variability in the dialects of RSL are investigated.

On the other hand, (Grenoble, 1992) states that there is definitely dialectal variations between Moscow and St. Petersburg. Differences are reported in approximately 50% of all the signs compared between Moscow and St. Petersburg.

Besides RSL, there exists the Signed Russian or “calqued sign speech” (Zaitseva, 2000), which directly interpolates grammar of spoken Russian, i.e. is a manually coded version of the Russian language. Signed Russian is used for formal and official situations and is regarded as the prestige, ‘academic’ variety of RSL by the death community in Russia.

The problem of correlation between of RSL and Signed Russian vocabularies has not received detailed coverage

in scientific works. However, (Zaitseva, 2000) points out that "the main gestures come to Signed Russian from RSL and form the largest and most stable lexical class of Signed Russian" (Zaitseva, 2000: 34). This is because the first language of a deaf person is RSL, and only after starting to attend school, the child begins to master Signed Russian. Gestures that are specific to Signed Russian include official style idioms, special vocabulary, and borrowing from foreign sign languages.

The most authoritative vocabulary of RSL is (Geilman, 1975), which actually manifests the norms of literary language for both RSL and Signed Russian.

4. Grammar and Phonology

The basic features of the RSL grammar are given, for example, in (Zaitseva, 2000; Davidenko and Komarova, 2006; Kimmelman, 2010; Lvovskaya, 2006; Prozorova, 2007; Shamaro, 2007; Kimmelman, 2014; Zavaritsky, 2015: 8–34). RSL nevertheless remains a low-resource language: despite recent interest in RSL, there is still a general lack of RSL descriptive grammars.

4.1 Phonology

In (Stokoe, 1960), gesture decomposition into five components was introduced: 1) handshape; 2) location; 3) orientation; 4) movement; 5) non-manual features (such as mimics). Specific forms and orientations of the hand, location and manner of movement are essential elements of any gesture, being, roughly speaking, analogous to distinctive phonological features of the sound languages. The set of realizations of these features is finite (see (Battison, 1978) for American sign language statistics). All phonological theories describing sign languages operate with these features in one way or another.

In the study (Klezovich, 2019) based on the annotation and analysis of more than 5000 images of RSL gestures there has been identified 116 configurations in the RSL, of which only 23 were phonemic. The most frequent configurations coincide with these of other sign languages (such as American, Israel, British, Dutch sign languages). Typologically, 23 is not a very extensive inventory.

Gestures in the sign languages of the world can divided into iconic and non-iconic. Iconicity is an essential feature of both sign and sound languages. Iconicity can be defined as a formal resemblance between the designee and the associated sign (sounds or handshapes, movements, locations) (Taub, 2012). (Kimmelman and Klezovich, 2018) introduces a project of iconicity patterns in sing languages, which takes into account RSL data.

4.2 Morphological meanings

Two types of morphological structures are distinguished in sign languages: linear and non-linear. Linear structures include affixation (joining a certain additional segment with a specific meaning to the main gesture), reduplication (full or partial repetition of a gesture), and word composition. RSL uses the most common strategy from the typological point of view: reduplication and world composition, to express some morphological meanings. Importance of affixation is low. The following morphological meanings can be found in the substantive domain of RSL:

a) natural gender marking: a range of RSL substantives are classified according to the category of the natural gender, indicating that the referent belongs to the male or female gender. To express the meaning of "male gender", the gesture is shown at the level of the upper part of the face (forehead); the meaning of "female gender" is transmitted by the localization of the gesture at the level of the lower part of the face (cheek).

b) plurality: in (Burkova, 2015), the following types of nominal plurality in RSL are distinguished: additive, collective, associative, and distributive. Each type of multiplicity is transmitted using specific lexical and morphological means. The "standard" plurality is additive, being expressed as the reduplication either of the entire gesture, or of its part, or non-manual components.

Another way of expressing additive multiplicity in RSL is making use a classifier constructions and quantifier gestures MANY, ALL, as well as numerals. In some cases, additive multiplicity is not expressed morphologically or lexically, being expressed by the context.

Collective multiplicity for contactless gestures is expressed morphologically, by modification of movement pattern, or analytically for contact gestures, using index gestures.

The meaning of associative plurality ("X and similar objects") in RSL is expressed by a combination of a significant gesture / several gestures that summarize the elements of the designated population and the word MISCELLANEOUS.

Distributive multiplicity (indicating a set of objects located at different points in space) for contactless gestures is expressed by a combination of reduplication and localization shift.

In the verbal domain of RSL time and aspect meanings can expressed morphologically (Davidenko and Komarova, 2006). The Past and Future are expressed analytically with words WAS and WILL BE, IT'S ABOUT TO, or lexically, with time adverbs (YESTERDAY, TOMORROW, THREE YEARS AGO etc.) Aspect meanings are marked with some change of movement character/pattern. For example, repetition corresponds to habituality, or multiplicativity, or distributivity, slow movement expresses durativity, a single sharp movement can be associated with completivity, or semelfactivity (i.e., completion or punctuality of a situation), etc. Special words such as READY/FINISHED (perfective), can be used as well.

4.3 Syntax

The main contribution in the study of RSL syntactical structures was made by (Kimmelman, 2012). It has

demonstrated, that two basic world orders exist in RSL: SVO and SOV. There are many factors that affect the world order in RSL. For example, classifier constructions are related to SOV world order, while verbal constructions are associated with SVO world order. Aspect markers, animacy and inanimacy of arguments, modifiers and objects belong to the world-order affecting factors as well.

5. Electronic Resources

Being an independent language with its own grammatical system and vocabulary, RSL has no writing system, and books, descriptive grammars, vocabularies and databases therefore are scarce. The only writing system developed for of RSL, is presented in (Dimskis, 2002). This system is based on the principles developed in (Stokoe, 1960), but did not gain much popularity in the deaf community of Russian Federation.

Among the principal electronic databases and electronic dictionaries, one should list "Thematic dictionary of Russian sign language" developed by the Moscow organization of All-Russian society of the deaf in 2006 (1480 signs), "Russian Sign Language Explanatory Dictionary RuSLED" (2537 videos with single words and phrases) with etymology of the signs (Voskresenskii et al., 2009); on-line video dictionary of Russian sign languages (Spreadthesign corpus), created in the framework of the European project "Spreadthesign" (14347 videos, comprising a mixture of words, phrases, utterances and variations); sign language corpus (RSLC, 2010-2011), recorded by the Novosibirsk State Technical University in 2010-2011 (230 spontaneous narratives by 43 native signers, annotated with ELAN tools). Surdoserver web-service (Surdoserver 2.0) and on-line dictionary (about 600 words and phrases). In the end of 2019, TheRuSLan database collected in SPIIRAS came into the world. TheRuSLan is aimed at RSL recognition tasks, being not very large, but recorded in 3D (Kagirov et al., 2020).

As (Kharlamenkov, 2017) states, most of the current RSL databases are either too small, or of poor quality: some of them are a mere mixture of lexical units that belong to different speech styles and dialects.

6. Conclusion

The main aspects of Russian sign language were sketched or mentioned in this paper: history, usage and dialects, phonology, grammar and databases. Despite a significant number of native speakers, RSL remains a low-resource language just for spoken communication of deaf people in Russia and neighbouring countries, and it is still poorly studied. The authors hope that the number of informative descriptions of RSL will increase in the future, and more databases will appear, which would enhance investigations of RSL in linguistic purposes and for applied systems, for example, automatic recognition, synthesis, and machine translation of Russian sign language.

7. Acknowledgements

The research is supported by the Ministry of Science and Higher Education of the Russian Federation, research project No. 14.616.21.0095, ID reference RFMEFI61618X0095.

8. Bibliographical References

Battison, R. (1978). Lexical Borrowing in American Sign Language. Silver Spring, MD: Linstok Press.

Bickford, A.J. (2005). The signed languages of the Eastern Europe. SIL Electronic Survey Reports.

Burkova, S., and Varinova, O. (2012). On the problem of regional and social variation in Russian Sign Language. In Fedorova O.V. (ed). *Russian sign language: the first linguistic conference*, Moscow, pages 127–143. (in Russian).

Burkova, S.I. (2012). Russian Sign Language: General Information (in Russian). Available at: http://rsl.nstu.ru/site/signlang

Davidenko, T.P., and Komarova, A.A. (2006). A short survey of RSL linguistics. In Komarova A.A. (ed.) *Current Issues in Sign Language*. Moscow, Russia, pp. 146–161.

Dimskis, L.S. (2002). Learning the sign language: A manual. Moscow: Academia (in Russian).

Geilman, I.F. (1975). Special tools for deaf communication. Dactylology and facial expression. Vol. 1–4, Leningrad, USSR. (in Russian).

Grenoble, L. (1992). An overview of Russian Sign Language. *Sign Language Studies*, 77:321–338.

Kagirov, I., Ryumin, D., Axyonov, A., and Karpov, A. (2020). Multimedia Database of Russian Sign Language Gestures in 3D. *Voprosy Jazykoznanija*, 1:104–123. (in Russian).

Kharlamenkov, A. (2017). Sign Language Vocabularies Review: https://surdocentr.ru/publikatsii/obzory-slovarej-zhestovykh-yazykov

Kimmelman, V. (2009). Reflexive pronouns in Russian Sign Language and Sign Language of the Netherlands: MA thesis in linguistics. Amsterdam: Universiteit van Amsterdam, 2009.

Kimmelman, V. (2012). Word order in RSL. An extended report. *Linguistics in Amsterdam*, 5(1):1–59.

Kimmelman, V. (2014). Information Structure in Russian Sign Language and Sign Language of the Netherlands. PhD dissertation, University of Amsterdam.

Kimmelman, V., Klezovich, A., and Moroz, G. (2018). IPSL: A Database of Iconicity Patterns in Sign Languages. Creation and Use. Proc. 11th International Conference on Language Resources and Evaluation (LREC'18), pages 4230–4234, Miyazaki, Japan. European Language Resource Association (ELRA).

Klezovich, A. (2019). Automatic Extraction of Handshapes Inventory in Russian Sign Language. In *NRU HSE. Series WP BRP "Linguistics"*, 86: In Print.

Lvovskaya, A. (2006). Analysis of Aspect and Tense System of RSL verb, based on the study by O. Dahl "Tense and Aspect Systems". Moscow.

Prozorova, E.V. (2007). Russian sign language as a subject of linguistic research. *Voprosy Jazykoznaniya*, 1:44–61. (in Russian).

Shamaro, E.Ju. (2007). Some facts concerning the time-aspect system of RSL. In Komarova A.A. (ed.) *Current Issues in Sign Language*. Moscow, pages 180–191.

Stokoe, W.C. (1960). Sign Language Structure: An Outline of the Visual Communication Systems of the American Deaf. Buffalo: Dept. of Anthropology and Linguistics, University of Buffalo.

Williams, H.G., and Fyodorova, P. (1993). The origins of the St. Petersburg institute for the deaf. In Fischer, R., and Lane, H. (eds). *Looking back: A reader on the history of Deaf communities and their sign languages. International studies on sign language and communication of the Deaf 20*. Hamburg: Signum-Verlag, 1993., pp. 295–305.

Zavaritsky, D.A. (2015) 100 phrases in Russian sign language. Phrasebook for clergymen. Moscow.

Zaytseva, G.L. (2000) Signed speech. Dactylology: Textbook for university students. Moscow. (in Russian)

Taub, S. F. (2012). Iconicity and metaphor. In Pfau, R. et al. (eds). *Sign Language: An International Handbook*. De Gruyter Mouton, Berlin, 2012, pp. 388–412.

Voskresenskii, A.L., Gulenko, I.E., and Khakhalin, G.K. (2009). The RuSLED dictionary as an instrument for semantic research. In Proc. «Dialog-2009» Annual International Conference, pages 64–68, Bekasovo, Russia (in Russian).

Russian Sign Language Corpus (RSLC), available at: http://rsl.nstu.ru/site/index/language/en

Spreadthesign corpus, available at: https://www.spreadthesign.com/en.gb/search

Surdoserver 2.0 service, available at: http://www.surdoserver.ru

Proceedings, pages 75–78
Paris, UNESCO Headquarters, 5-6 December, 2019.

Language Technology Research at MILE Laboratory

Ramakrishnan A G and Madhavaraj A
MILE Laboratory, Department of Electrical Engineering
Indian Institute of Science, Bangalore, India
(agr, madhavaraja)@iisc.ac.in

Abstract

Medical Intelligence and Language Engineering Laboratory at the Indian Institute of Science developed Kannada and Tamil OCR and TTS systems and deploying them, created Braille and audio books for blind students and won Manthan Awards. Algorithms proposed for scene and born-digital word image recognition consistently retained the top positions in ICDAR Robust Reading Competitions since 2011. The TTS systems were adjudged second best in Blizzard TTS Challenge in 2013 and 2014. Current research is on solving the problems of unlimited vocabulary in building ASR systems for Dravidian languages of Kannada and Tamil, by using sub-word units for language modeling.

Keywords: OCR, TTS, ASR, transcription, Tamil, Kannada, Braille books, audio books, Manthan award, script recognition.

Résumé

இந்திய அறிவியல் கழகத்தின் மருத்துவ நுண்ணறிவு மற்றும் மொழிப் பொறியியல் (மநுமொழி) ஆய்வகம், கன்னடம் மற்றும் தமிழ் ஓ.சி.ஆர் மற்றும் டி.டி.எஸ் மென்பொருள்களை உருவாக்கி அவற்றின் மூலம் பார்வையற்ற மாணவர்களுக்கு பிரெய்ல் மற்றும் பேசும் புத்தகங்களை உருவாக்கி, மந்தன் விருதுகளை வென்றது. செல்பேசியில் பிடிக்கப்பட்ட மற்றும் பிறப்பு-டிஜிட்டல் சொல் படங்களை அடையாளம் கண்டு கொள்வதற்காக முன்மொழியப்பட்ட மநுமொழி வழிமுறைகள் 2011 முதல் ஐசிடிஏஆர் வலுவான வாசிப்பு போட்டிகளில் தொடர்ந்து முதலிடங்களைத் தக்க வைத்துக் கொண்டன. 2013 மற்றும் 2014 ஆம் ஆண்டுகளில் ப்ளிஸ்ஸார்டு டிடிஎஸ் சவாலில் மநுமொழி டிடிஎஸ் மென்பொருள்கள் இரண்டாவது சிறந்த இடத்தைப் பிடித்தன. தற்போதைய ஆராய்ச்சி, மொழி மாடலிங் செய்வதற்கு துணை சொல் அலகுகளைப் பயன்படுத்துவதன் மூலம் கன்னட மற்றும் தமிழ் ஆகிய திராவிட மொழிகளுக்கு ஏ.எஸ்.ஆர் அமைப்புகளை உருவாக்குவதில், வரம்பற்ற சொற்களஞ்சியத்தின் சிக்கல்களைத் தீர்ப்பதாகும்.

1. Introduction

The MILE laboratory at the Department of Electrical Engineering has made significant contributions to analysis of speech signals, text-to-speech conversion systems for Indian languages, document and scene image analysis and recognition and online handwriting recognition in Indian languages and is currently working on automated speech recognition for Tamil, Kannada and Hindi languages. The research focuses on real life applications of signal, image processing and pattern recognition in solving crucial problems in language technology, of particular relevance to Indian scenario, which is followed up with actual technology development. The work on document image analysis, script recognition, OCR of printed text, scene text recognition and text-to-speech conversion have been motivated from the commitment to develop deployable technology to enable visually challenged people to be able to have access to (by machine reading of) any printed book in Indian languages. In addition, free tools have been developed to read web text in any major Indian language in one's own script. Also, Indic kBD, a tool for typing in any Indian script on Linux & Windows using QWERTY keyboard using anyone of many keyboard mappings has been created and made available for download.

2. Speech Analysis, Synthesis & Recognition

2.1 Time-Domain Features for Speech Analysis

Novel techniques have been proposed based on knowledge-based acoustic-phonetic approach to detect stop closure-burst transitions and epochs in speech. The new nonlinear feature, defined in the time-domain, called plosion index, is robust in detecting stop closure-burst transitions and performs much better than complex feature vectors of a large dimension. Extension of this, called the dynamic plosion index, has been shown to be robust in detecting instants of significant excitations in voiced speech and also the QRS complexes of noisy ECG.

2.2 DCT based Pitch-Synchronous Pitch Modification for Prosody Modification

A novel algorithm has been proposed for pitch modification. The linear prediction (LP) residual is obtained from pitch synchronous frames. The dimension of the DCT coefficients of the residual is modified by truncating or zero padding, and then the inverse DCT is obtained. This period-modified residual is then forward filtered to obtain the pitch modified speech. The radii of the poles of the filter are modified to smoothen the LP spectrum. This minimizes the mismatch between the pitch modified signal and the LP spectrum due to the change in the positions of the pitch harmonics. The technique has been applied to create interrogative sentences from affirmative ones in our Tamil TTS.

2.3 TTS Systems for Kannada and Tamil

The concatenative TTS system uses a basic unit, that is distinctly different from all the TTS systems in the world. This is very close to what is known as akshara in Indian languages and the concatenation is performed only across similar vowels, which makes it smooth and glitch-free. The *Thirukkural* Tamil TTS system developed is being used by over 1000 blind student members of Anna Centenary Library, Chennai. The *Madhura Vaachaka* Kannada TTS system has been used to convert high school and PUC books to audio books by www.kannadapustaka.org.

3. Document Image Reconstruction, Analysis, and Recognition

3.1 Knowledge-Driven Deep Models for Superresolution of Low-Resolution, Scanned Binary Document Images

While the widely accepted view in computer vision today is to use end-to-end approaches using deep neural networks (DNN), this work has convincingly shown that the performance of existing state-of-the-art DNN models for super resolution can be significantly improving by suitably modifying the objective function, driven clearly by the knowledge of the specific vision problem in question. Computationally efficient superresolution models are obtained by nonlinear fusion of the outputs of well-known image interpolation techniques. This method has been used to significantly improve the resolution of binary document images so that human readability as well as OCR recognition accuracy improve appreciably. This has been filed as an Indian patent and also a PCT application.

3.2 Script Recognition at the Word Level in Multilingual Documents

An algorithm has been proposed to identify the script of each word in a multiscript document image. Gabor and DCT features were independently evaluated for their effectiveness using different classifiers. Gabor features with support vector machine classifier has given promising results; i.e., over 98% for bi-script and tri-script cases.

3.3 Kannada OCR with Performance Better Than Google's Tesseract OCR

Inspired by the rich feedback in the ascending visual pathway in higher mammals, attention -feedback has been proposed for improving the performance of printed text and handwriting recognition systems. Different types of recognition errors are identified at the different stages of the machine learning system, and this feedback is used effectively to revise the binarization, line segmentation, character segmentation and recognition in printed text. This innovative idea has been filed as a patent. This has resulted in a Kannada OCR (Lipi Gnani, which has been shown to perform better than the latest version of Tesseract OCR on 250 benchmarking images (Shiva Kumar and Ramakrishnan, 2020). Tesseract OCR is being developed for 3 decades, originally by HP Labs for a decade, and then taken over by Google.

3.4 Analysis and Recognition of Camera-Captured Document Images

Techniques have been proposed to binarize coloured documents captured by cameras. New approaches were proposed for recognition of script at the level of the word in multi-script documents and for text extraction from complex, colour document images. An edge-based connected component approach has been proposed for binarization of color documents. It handles documents with multi-colored texts with different background shades; deals with text of widely varying sizes, not handled by local binarization methods; automatically computes the binarization threshold without requiring any input parameter.

4. Development of ASR Systems for Three Indian Languages

In this section, we elaborate the development of ASR systems for three Indian languages namely Hindi, Kannada and Tamil. In order to develop a good automated speech recognition (ASR) system, we require (i) high quality transcribed speech data in the order of several hundred hours and (ii) multi-domain text corpus containing several million words. Conventional ASR systems use graphical models like finite state transducers (FST) and stochastic and neural models like hidden Markov model (HMM) and deep neural networks (DNN) (Hinton et al., 2012). Recent end-to-end connectionist temporal classification (CTC) based techniques have been successfully applied to build large-scale ASR systems (Amodei et al., 2016). However, the size of the speech corpus required for the CTC models is much larger than that needed by the graphical models.

Due to limited data resources, we have used FST based models, which have been trained using 137, 280 and 180 hours of speech for Hindi, Kannada and Tamil, respectively. Our ASR systems have been built using the Kaldi open source toolkit (Povey et al., 2011). The building blocks of our ASR system are explained below.

4.1 Speech Data Collection and Correction

Since we require a large amount of transcribed speech corpus for training the acoustic model of the ASR, we have developed a speech recording tool that loads text prompts from our database and the volunteer reads the prompts one by one. Speech data has been collected using Sennheiser PC-8, Plantronics C320-M headphones and mobile handsets. The tool has provisions to rectify any errors in the recorded text/speech at the time of recording. Natural language processing tools for converting numbers, symbols and abbreviations have been integrated into the tool so that minimal manual effort is required to make the data ASR-ready. We have also developed an online transcript correction tool so that any errors uncorrected by the speaker while recording can be corrected at a later stage. After this, the transcribed speech data would be used by our ASR for training. Measures have been taken to ensure that the phone distribution in the spoken utterances matches that of the text corpus. Using this tool, we have collected speech from around 2500 native speakers of Hindi, Kannada and Tamil.

We have also collected a large amount of text corpus in these languages from Wikipedia articles, newspapers, magazines and books. The collected text has been pruned by removing Unicode errors and converting numerals, abbreviations and symbols. This is used to build the language model.

4.2 Design of the ASR Systems

The lexicon/pronunciation model has been created by getting all unique words from the text corpus and performing a grapheme to phoneme conversion. Schwa deletion in Hindi (Deepa et al., 2004) and voiced/unvoiced phonation rules for stop consonants in Tamil have been incorporated for better phone modeling. The lexicon model can be thought of as a map from word to phone sequences. Alternate pronunciations have been included for relevant

words in the lexicon with appropriate pronunciation probabilities (Chen et al., 2015).

Using the transcribed speech, we train a series of models: monophone, triphone and DNN to get the speech transcribed at the phone-level (Madhavaraj and Ramakrishnan, 2017). Finally, the DNN model is used with the word level trigram language model for decoding. Viterbi decoding with beam search is used during testing to get the best possible sequence of words from the given speech. The size of the vocabulary of train and test data and word error rate performance of our ASR systems are given in Table 1.

Language	Training data	Test data	vocabulary size	WER
Hindi	137	45	65421	9.51
Kannada	280	67	200690	11.45
Tamil	180	54	189644	13.56

Table 1: Training and test data sizes (in hours), vocabulary size (in words) and word error rate (in %) of MILE ASR systems for Tamil, Kannada and Hindi.

5. Performance Enhancement of ASR Systems

Traditional ASR systems use mel frequency cepstral coefficients (MFCC) as speech features for acoustic modeling. These features are inspired by the human auditory mechanism and contain information about speaker identity, phone identity, stress, emotion, age and gender. However, for speech recognition, features are desired containing information only about phone identity. We have proposed two techniques: the first is a hybrid feature/model engineering technique based on scattering transform, and the second adapts the DNN model to suppress speaker variability. Both techniques succeed in extracting phonetic information as evident from the reduction in WER (see Table 2). These techniques are illustrated in the following two subsections.

Features/model	WER
Baseline MFCC	13.56
LFBE	13.48
Scattering transform (order 1)	13.16
Scattering transform (order 2)	12.36

Table 2: Comparison of word error rates (in %) of MILE Tamil ASR for different architectures and features.

5.1 Scattering Transform based Features for Better Acoustic Modeling

In this experiment, we have proposed a new DNN architecture employing a cascade of 1-D and 2-D filterbank layers which are essentially 1-D and 2-D convolution layers initialized with Gabor filter coefficients with various center frequencies and orientations. This architecture is motivated by filterbank learning techniques from raw speech waveform (Sainath, 2015) and uses scattering spectrum as front-end features (SainathScattering}. The features obtained from 1-D and 2-D filterbank layers are combined and fed to a 7-layer feed-forward DNN for predicting the phoneme labels. This architecture models the acoustic features better, since it learns the features directly from the raw waveform. Using these filterbanks for Tamil ASR, we get a relative WER reduction of 2.94\% and 8.85\%, respectively, compared to the baseline features as shown in Table 2. More details about this experiment can be found in (Madhavaraj and Ramakrishnan, 2019).

5.2 Speaker Adaptation using DNN Co-activation Modeling

This involves suppressing speaker-specific information contained in the speech signal and extracting features relevant only for phone identification. Here, we propose a supervised speaker adaptation technique for DNN, which estimates prior statistics of node activations in every DNN layer from the training data and adapts the weights based on the activations obtained from the adaptation data. The DNN weight update optimizes a loss function which combines cross-entropy loss and KL-divergence measure between the prior activation statistic and adaptation data's statistic. Just by modifying the loss function, we obtain an absolute WER reduction of 2.44\% over the baseline model. The results of our experiments with other variants of this training strategy are listed in Table 3.

Model adaptation type	WER
Baseline architecture	13.56
Mean normalization at every DNN layer	13.48
Mean & variance normalization at every DNN layer	13.44
Mean normalization at the first affine layer of DNN	11.62
Mean & var. normalization at the first affine layer of DNN	11.12

Table 3: Comparison of performance of Tamil ASR for different speaker adaptation schemes for DNN-based acoustic models.

6. Extending the Vocabulary of ASR Systems using Subword Modeling

Handling the infinite vocabulary problem is a major task in improving the recognition accuracy of ASR systems for Tamil and Kannada. This problem arises due to morphology, inflexion and agglutination properties of the languages. Graphical model based ASRs require a finite set of words, and it is impossible to contain these languages within a finite vocabulary and build the system. To tackle this issue, we propose subword modeling, where the vocabulary contains only the subword prefixes, infixes and suffixes with proper identification markers for the language model to learn the order of subwords. The subword modeling experiments conducted for Tamil ASR are explained below.

6.1 Word Morphology based Language Modeling

In this experiment, we use the Morfessor toolkit (Smit et al., 2014) for subword modeling. Morfessor is a statistical

machine learning tool used for morphological analysis to segment a given word. These subwords are used as basic units in our lexicon for recognition. We learn the language model by converting the training corpus containing words into a sequence of subword tokens. The lexicon preparation for the subword dictionary is a straightforward task for Indian languages, since they have an almost one-to-one correspondence between graphemes and phonemes. The rest of the ASR systems are built as explained in (Madhavaraj and Ramakrishnan, 2017). During post-processing, the subwords are joined into words using the identification markers in the recognized text. We obtain a WER comparable to that of the word-level ASR as reported in Table IV.

6.2 Maximum Likelihood based Language Modeling

This uses byte pair encoding (Sennrich et al., 2016) to perform subword modeling, where the list of subwords, their occurrence and co-occurrence probabilities we are derived and based on a specifically designed subword finite state transducer (FST), the most likely segmentation for a given word is estimated. Since segmenting a given word into subwords is a combinatorial explosion problem, maximum likelihood (ML) estimation is employed through expectation-maximization procedure and this problem is posed as a weighted-FST (WFST) graph search problem. Two different methods are proposed for ML estimation, namely Viterbi and forward-backward techniques, whose performances are listed in Table 4.

6.3 Manual Modeling

The performances of both morphology-based and ML-based techniques depend highly on the quality and diverseness of the corpus. Yet, there is a high probability that some of the subwords derived by these techniques may not be valid prefixes, infixes or suffixes. So, we manually construct the lexicon graph for Tamil words for different nouns, pronouns, adjectives, adverbs, numbers, verbs and infinitives. The ASR now uses a lexicon of only hand-labeled subword units for recognition. We have also created a provision to add new words into this lexicon graph which cannot be analyzed morphologically. One advantage with this modeling is that the lexicon graph can be readily used for many NLP tasks such as part of speech tagging, lemmatization and text translation.

7. Conclusion

The commitments and contributions of MILE laboratory over the past two decades in performing fundamental research in technologies for Indian languages have been described briefly. All the data we use have been collected by us: India has a huge population and so, there is no dearth for creation of standard databases.

8. Acknowledgements

The first author gratefully acknowledges Tata Trust Travel Grant for funding him to travel and participate in this conference. Immense thanks are also due to the Technology Development for Indian Languages (TDIL), Ministry of Information Technology, Government of India, for funding many of his projects in language technology, which has made it possible for him to be invited to this conference. Thanks are also due to many students and research staff, who enriched his knowledge and experience.

9. Bibliographical References

Geoffrey Hinton, li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Phuongtrang Nguyen, Tara Sainath, and Brian Kingsbury. (2012). Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Processing Magazine*, vol. 29, pp. 82–97.

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen. (2016). Deep speech 2: End-to-end speech recognition in English and Mandarin. Proc. 33rd International Conf. on Machine Learning (ICML) Vol. 48, p. 173–182.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. (2011). The Kaldi speech recognition toolkit. Proc.IEEE Workshop Automatic Speech Recog. Understanding.

S.R. Deepa, Kalika Bali, A.G. Ramakrishnan, and Partha Pratim Talukdar. (2004). Automatic generation of compound word lexicon for Hindi speech synthesis. Proc. Fourth International Conf. on Language Resources and Evaluation (LREC'04), May, ELRA.

Guoguo Chen, Hainan Xu, Minhua Wu, Daniel Povey, and Sanjeev Khudanpur. (2015). Pronunciation and silence probability modeling for ASR," Proc. 16th Interspeech.

A. Madhavaraj and A. G. Ramakrishnan (2017). Design and development of a large vocabulary, continuous speech recognition system for Tamil. Proc. 14th IEEE INDICON, Dec, pp. 1–5.

Tara N. Sainath, Ron J. Weiss, Andrew W. Senior, Kevin W. Wilson, and Oriol Vinyals. (2015). Learning the speech front-end with raw waveform CLDNNs. Proc. 16th Interspeech.

V. Peddinti, T. Sainath, S. Maymon, B. Ramabhadran, D. Nahamoo, and V. Goel. (2014). Deep scattering spectrum with deep neural networks. Proc IEEE ICASSP, pp. 210–214.

A. Madhavaraj and A. G. Ramakrishnan. (2019). Scattering transform inspired filterbank learning from raw speech for better acoustic modeling. Proc. IEEE Region 10 Conf. (TENCON), Oct, pp. 1154–1158.

Peter Smit, Sami Virpioja, Stig-Arne Gronroos, and Mikko Kurimo. (2014). Morfessor 2.0: Toolkit for statistical morphological segmentation. Proc. Demonstrations 14th Conf. of European Chapter of the Association for Computational Linguistics. Apr., pp. 21–24, ACL.

Rico Sennrich, Barry Haddow, and Alexandra Birch. (2016). Neural machine translation of rare words with subword units. Proc. 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, Aug., pp. 1715–1725, ACL.

Shiva Kumar H R and A G Ramakrishnan. (2020). Lipi Gnani - A versatile OCR for documents in any language printed in Kannada script. ACM Transactions on Asian and Low-Resource Language Info Processing (TALLIP).

Proceedings, pages 79–82
Paris, UNESCO Headquarters, 5-6 December, 2019.

How aspects of descriptive and formal linguistics can inform LT for all languages

Lars Hellan
NTNU
Nidareid 5, 7017 Trondheim
lars.hellan@ntnu.no

Abstract

What exists for many 'digitally less resourced' languages (LRL) are grammars and dictionaries. Common to formats like CSV (underlying many dictionary tools) and grammar encoding formalisms like attribute-value matrices and feature unification, is that through 30-40 years of development and sustenance of life-long projects, sustainable tools for 'whole language'-size resource creation are by now well established. Interesting possibilities resides in exploring ways in which the content of such resources can be channeled onto further types of structures and processing tools. The paper exemplifies this point through procedures for digital lexicon creation and annotation schemata reflecting advanced formal analysis.

Keywords: Less Resourced Languages, Valence dictionaries, Toolbox, Typed feature grammar, Corpus annotation schemata

Résumé

Hva mange 'digitalt lite utviklede språk' har er grammatikker og ordbøker. Felles for formater som CSV (bruket i mange ordboks-redsakaper) og grammatikkformalismer som AVM of faktor-unifikasjon, er at det nå foreligger mange bærekraftige 'hel-språk'-ressurser. Det ligger interessante muligheter i i å utforske muligheter for hvordan innholdet i slike ressurser kan bli kanalisert inn i videre typer prossesseringsredsakaper. Artikkelen eksemplifiserer dette gjennom prosedyrer for fremstilling av digitale leksika og annotasjonsskjema som avspeiler avansert formell analyse.

1. Introduction

An aspect of language technology (LT) somewhat less highlighted in recent years is its contributions to grammar description and lexicography, in respects as basic as those of formal structuring of content and enhancing consistency. Albeit the creation of grammars and dictionaries are tasks belonging within linguistics/lexicography, when these activities are conducted using computationally tractable formats, the consistency thereby attained can be readily further exploited in the creation of processing tools and applications.

Traditionally, the creation of grammars and dictionaries is very much 'rule based' – although the induction of such resources from corpora and through machine learning are exciting issues in current NLP, hardly any language has had its basic grammar and lexicon resources produced by such means. For most 'digitally less resourced' languages (LRL), procedures in this domain are not even in question given the scarcity of digital resources. What nevertheless exists for many such languages are grammars and dictionaries – in many cases just in printed versions (or even just handwritten), but since the 1990ies also to a growing extent with dictionaries encoded in tools like Shoebox and Toolbox. Formal grammars of LRLs are much rarer, but common to formats like CSV (underlying dictionary tools as mentioned) and grammar encoding formalisms like attribute-value matrices and feature unification, is that through 30-40 years of development and sustenance of (person-) life-long projects, sustainable tools for 'whole language'-size resource creation are by now well established. From the perspectives of 'LT for all' related to indigenous languages, most of which will count as LRL, interesting possibilities resides in (i) furthering the use of formats like those mentioned to ever more languages; (ii) exploring ways in which the content of such resources can be channeled onto further types of structures and processing tools.

It is well known that Toolbox dictionaries can be converted into lexicon modules of feature structure grammars; we describe how a further step in this line of development underlies the creation of a large scale valence dictionary of the West African language Ga (Kwa), viz. Dakubu 2010, 2011.

The representation of grammatical structures of types prevalent in many LRLs, although less common in European languages, such as complex verb extensions in Bantu and Serial Verb Constructions (SVCs) in Kwa, is by now well established using feature structure formalisms. We exemplify formalisms for sorting of lexical items and for corpus annotation which reflect feature structure analysis, thereby opening for further connections between formal analysis in the domain of lexicon and grammar on the one hand, and notational features applicable in dictionaries and corpora.

In both cases what is exemplified are ways in which research in, and resource development for, LRLs can be enhanced using well established formalisms within logic and computational linguistics, based on independently established linguistic work. These may not yet yield digital applications ready for a user marked, but with attention thus directed to types of resources that already exist or are within reach for a multitude of LRLs, one may reduce the distance between linguistic resources and LT.

2. Creating a valence dictionary from a Toolbox lexicon

The digital lexicon (Dakubu 2010) is an amended version of a Ga Toolbox lexicon project holding data for a general-purpose dictionary. (Dakubu 2011) is a free-standing linguistic monograph. The former consists of 80,000 lines of code with 7080 entries, of which 5014 are for nouns and 935 for verbs. Here valence codes are written into the lexical entry following the general field style of Toolbox, where for the item *ba*, for instance, fields named \pdl-\pdv represent inflectional information of the lexeme , and the fields \xe, \xg, \xv together constitute a standard linguistic glossing with \xv as a word-and-morph break-up, \xg as morphological and English gloss, and \xe as a free English translation. With valence information added, a verb with more than one valence frame has one entry specified per frame; thus the verb *ba* 'come', for instance, is represented by 18 different entries in this edition of the Toolbox file. In this way, 547 verb lexemes from the original file are represented through altogether 2006 entries. The valence specification follows principles and formalization laid out in (Hellan and Dakubu 2010),[1] the *Construction Labeling* (*CL*) system. In this formalization one of these frames can be represented on the form given in (1), to be read as 'a verb-headed intransitive syntactic frame where the subject carries an agent role and the situation expressed belongs to the type 'MOTIONDIRECTED'.

(1) v-intr-suAg-MOTIONDIRECTED

The semantic specification here consists of two parts, *semantic role* as exemplified by 'suAg' and *situation type* as exemplified by 'MOTIONDIRECTED', the latter out of a total inventory of about 130 situation types.[2]

The classification using all the parameters recognizes about 100 construction types, which for mono-verbal constructions could also be seen as valence types. This addition to the Toolbox file thus constitutes a valence lexicon, with illustrating sentences. A small corpus further illustrates these construction types.[3]

With a set of 2000 entries classified by strings like (1), the valence notation allows one to investigate the frequency of frames used relative to these frames, correspondences between syntactic and semantic structure, the clustering of certain valence types for sets of verbs, and more.

The specifications of the Toolbox valence lexicon are also used for classification in a Ga lexicon with the 2000 verb entries at the online (4-language) valence lexicon *MultiVal*.[4] The lexicon is also used in a computational grammar of Ga based on the HPSG framework. These are examples of how information, once digitally encoded, can be recast in other formats and used for other purposes.

This briefly illustrates how a general-purpose lexicon can be expanded to a velence lexicon, in turn used in a so-called 'deep processing' grammar and in an online multilingual valence resource. All steps are technically straightforward, only the task of specifying valences is time-consuming, and can only be done by a linguist interested in creating such specifications – which is after all the normal way of creating linguistic resources.

[1] Also see (Dakubu and Hellan 2017).
[2] See (Dakubu 2011) and (Hellan and Dakubu 2010).
[3] See https://typecraft.org/tc2wiki/Ga_Valence_Profile. The data are searchable, so that a search for, e.g., the constructional factor *obPostp* ('object is a postposition') yields an array of urls for the sentences instantiating the factor.
[4] Cf. (Hellan and Beermann. 2014).

3. Representing grammatical analysis in corpus annotation

In this section we illustrate a schema for integrating grammatical analysis into corpus annotation. The schema provides construction-level annotation tags which in one-line strings provide much of the information that could otherwise be expressed in multi-tier syntactic and semantic annotation. The strings are subject to semi-automatic consistency control, and can also be applied in valence specification in lexicons, grammatical parsing, and more. The tag system is referred to as Construction Labeling (CL), mentioned in the previous section, earlier presented in Hellan and Dakubu 2010 and Dakubu and Hellan 2017, but with the added capacity of serving as types in a Typed Feature Structure (TFS) system, enabling the consistency control and the parsing functionality.

To illustrate the complexity of information that can be accommodated, we use examples from Bantu instantiating verbal derivation and what may be called 'skewed' semantics.

The construction tags can be combined with standard word-by-word&morph-by-morph IGT annotation, as in TypeCraft (cf. Beermann and Mihaylov 2014), adding just a single line as annotation to the verb, as schematically illustrated with an example of verb derivation from Citumbuka (Bantu).

(2)

Mary wa-ka-mu-phik-**isk**-a		John	nchunga
Mary 1SM-Pst-1OM-cook-Caus-FV		John	beans
N	V	N	N
	vCaus-dbobCs		

'Mary made John cook beans'

vCaus means that the head is a verb and has a causative morpheme, and *dbobCs* means that the construction is a double object construction 'derived' through causativization and with the corresponding semantics non-isomorphic to the syntactic structure, a constellation we refer to as 'skewed' semantics.

We can make specifications of arguments of a derived verb in terms of their derivational histories, e.g., extending the formula *vCaus-dbobCs* to

(3) *vCaus-dbobCs-suC-obCsu-ob2Cob*

where the added items read as follows, similar to a formalism used in Relational Grammar:

suC - subject created by Causativization
obCsu - object derived ('demoted') from subject by Causativization
ob2Cob - second object derived from object by Causativization

Expanding from what was said in section 2, each CL tag is a string consisting of, first, a label specifying POS of head of the construction and salient morphological marking (like *vCaus* in (2)), second, a label designating the overall structure of a construction (encoding notions like intransitive, transitive, ditransitive/double object, etc. (such as *dbobCs* in (2)), third a string of labels classifying features of the arguments - first syntactic features and then semantic features -, and finally a string of labels for TAM features and situational content. (4) further illustrates the format, applicable to a sentence like *John ate the cake*:

(4) *v-tr-suAg_obAffincrem-COMPLETED*

Whenever a putative CL string is composed, the labels of the string have to match – for instance, if one label is *intr*, then there cannot be an argument label prefixed by *ob*, since *intr* is not defined such a label. A processing mechanism enforcing such consistency is provided using a Typed Feature Structure (TFS) system, in which the CL tag labels are defined as *types*.

Information in such a system is generally exposed through Attribute Value Matrices (AVMs), where each AVM belongs to a type, and attributes are introduced (declared) according to the following conventions:

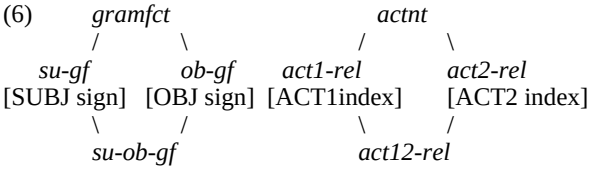
(5) *[A]* A given type introduces the same attribute(s) no matter in which environment it is used.

[B] A given attribute is declared by one type only (but occurs with all of its subtypes).

In a TFS representing a grammar, there are many type hierarchies, representing POS, tenses, semantic roles, etc.; some of these hierarchies do without attributes, while the following ones do.

Types for grammatical functions (values of 'GF') and actants (values of 'ACTNT') include those indicated below, the *gramfct* subtypes declaring GF attributes ('SUBJ' and 'OBJ') and the *actnt* subtypes declaring semantic participant attributes ('ACT1' and 'ACT2'):

(6)



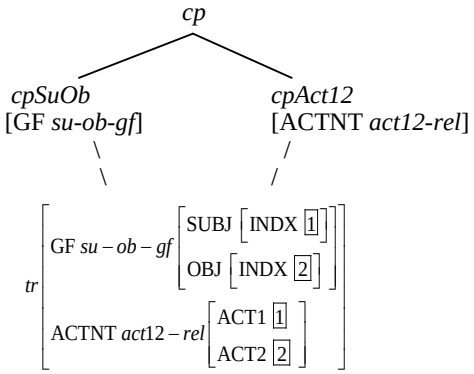
With such features as basis, one can represent, e.g., (2) as (7), which is an AVM representing a construction, which involves a specification of grammatical functions and actants acting together, identified through the attributes GF and ACTNT, neither of which are introduced in (6), but which are introdued at the level of constructions.

(7) AVM for double object construction with causative semantics and causative derivation:

$$dbobCs\left[\begin{array}{l}\text{HEAD } verb \\ \text{GF}\left[\begin{array}{l}\text{SUBJ}\left[\text{INDX } \boxed{1}\right] \\ \text{OBJ}\left[\text{INDX } \boxed{2}\right] \\ \text{OBJ2}\left[\text{INDX } \boxed{3}\right]\end{array}\right] \\ \text{INDX } \boxed{4} \\ \text{ACTNT}\left[\begin{array}{l}\text{PRED } 'cause' \\ \text{ACT0 } \boxed{4} \\ \text{ACT1 } \boxed{1} \\ \text{ACT2 } \boxed{6}\left[\begin{array}{l}\text{PRED } 'cook' \\ \text{ACT1 } \boxed{2} \\ \text{ACT2 } \boxed{3}\end{array}\right]\end{array}\right] \\ \text{D-BASE } sign\left[\begin{array}{l}\text{GF}\left[\begin{array}{l}\text{SUBJ}\left[\text{INDX } \boxed{2}\right] \\ \text{OBJ}\left[\text{INDX } \boxed{3}\right]\end{array}\right] \\ \text{ACTNT } \boxed{6}\end{array}\right]\end{array}\right]$$

In their capacity as types, CL labels define AVMs at the formal level of constructions, thus as subtypes of *cp* with all the attributes defined for *cp*. Type definitions sustaining the type *tr* ('transitive') reflect this, as shown in the following. The definition of *tr* is achieved through a join of *cp* specifications:

(8)



tr is thus formally defined as a type of *sign*, or *construction*. Similar depths of specification are required for all CL labels.

When CL labels occur in a string, as in (4) , they unify. To illustrate, the types to which the labels in (4) correspond are indicated in (9), and the unification result is (10):

(10)

v - - - [HEAD *verb*]

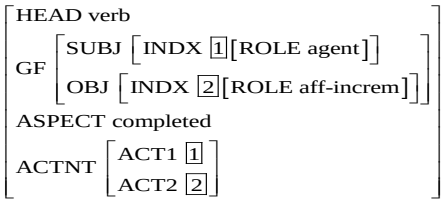
tr - - - *(cf. (9)*

suAg - - - [GF [SUBJ [INDXROLE *agent*]]]

obAffincrem - - - [GF [OBJ [INDXROLE *aff – increm*]]]

COMPLETED - - - [ASPECT *completed*]

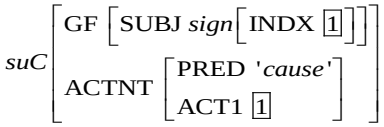
(11)



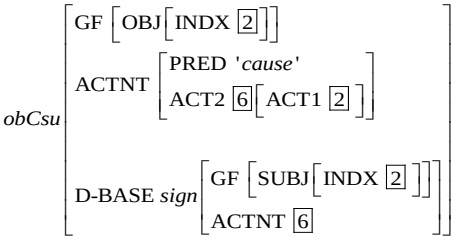
Returning to the more complex label in (3), the AVMs for the 'derivational histories' will be as in (12), the unification of which with a structure for dbobCs in isolation will be the structure in (8);

(12)

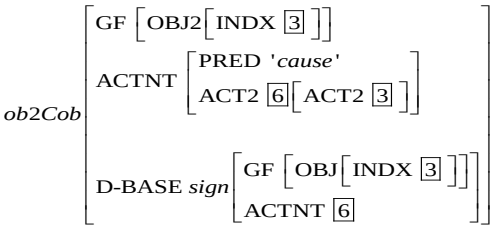
a.



b.



c.



Unification presupposing feature compatibility, a control of consistency is inbuilt in this formalism.

It is possible to run the TFS system as a parser where the IGT annotations serve as a 'pre-processed' input, and where the CL type assigned to the verb acts as a valence requirement. Consistency relative to the whole IGT can then be similarly imposed.

With such a parsing facility one can also generate for each sentence a detailed total structure, exposing for instance which parts of a sentence are subject and object, information one otherwise will expect to find in a treebank.

A design as now outlined addresses one aspect of what concerns correctness of annotation – that of consistency. Another aspect is of course factual correctness. In-between lurks the issue of 'using a correct annotation set'. Within a given project building up a database, it can be essential that the same tags are used for the same phenomena. But in a general perspective, there are many reasonable ways in which to name a phenomenon and assign tags to it.

In the present approach, one can freely add labels to the defined set, as long as they have concise definitions into the TFS system. Thus, if it is a matter of an alternative tag for an already 'tagged' phenomenon, one just equals the tags. In cases of a not yet accommodated phenomenon, more will be involved, ranging from filling a gap in an already established paradigm, to creating a new analysis in the TFS, the latter of course interesting but also more involved.

4. Conclusion

The CL annotation tagset consists of symbols which are on the one hand descriptive labels, and on the other hand labels for types reflecting multiple layers of analysis. The descriptive labels allow one to stay within the IGT over-all format, while their type definitions allow for additional layers of analytic representations, and for the possibility of defining semi-automatic consistency-checking procedures. We have illustrated its relevance both in the development of computational lexical resources from linguistic lexicons, and in the design of a corpus annotation schema reflecting 'deep' analytic features.

The general point that we want to make through this illustration is that descriptive linguistic resources need not be far from constituting interesting digital resources, also for so-called less resourced languages, and that 'deep' and formal structures of languages can be readily reflected in annotation schemata applied to small or larger corpora, equally readily for less resourced languages with a decent descriptive literature, as for well resourced languages.

5. References

Beermann, Dorothee and Mihaylov, Pavel. 2014. Collaborative databasing and Resource sharing for Linguists. *Languages Resources and Evaluation* 48. Dordrecht: Springer, 1-23.

Dakubu, Mary Esther Kropp. 2010. 'Ga_verb_dictionary_for_digital_processing'. https://typecraft.org/tc2wiki/Ga_Valence_Profile

Dakubu, Mary Esther Kropp. 2011. Ga Verbs and their constructions. Monograph ms, Univ. of Ghana.

Dakubu, M.E. Kropp and Lars Hellan. 2017. A labeling system for valency: linguistic coverage and applications. In Hellan, L., Malchukov, A., and Cennamo, M (eds) *Contrastive studies in Valency*. Amsterdam & Philadelphia: John Benjamins Publ. Co.

Hellan, Lars and M.E. Kropp Dakubu. 2010. *Identifying verb constructions cross-linguistically*. In *Studies in the Languages of the Volta Basin* 6.3. Legon: Linguistics Department, University of Ghana.

Hellan, Lars, and Dorothee Beermann. 2014. Inducing grammars from IGT. In Z. Vetulani and J. Mariani (eds.) *Human Language Technologies as a Challenge for Computer Science and Linguistics.* Springer.

Proceedings, pages 83–86
Paris, UNESCO Headquarters, 5-6 December, 2019.

Multilingual Crowdsourcing Methodology for Developing Resources for Under-resourced Indian Languages

Karunesh Kr. Arora[1], Sunita Arora[1], Mukund Kumar Roy[1], Shyam S. Agrawal[2]
CDAC[1], KIIT[2]
C-56/1, Sector 62, Noida, India[1], Sohna Road, Gurugram, India[2]
{karunesharora, sunitaarora, mukundkumarroy}@cdac.in, ss_agrawal@hotmail.com

Abstract

Huge Data collection challenge gets intensified for Under Resources Languages especially for large variety of Indian languages. We propose building of a common framework for collecting, monitoring and evaluating speech resources irrespective of the language. Common phone set for transcribing and annotating the data, easy portability, configuring different languages, indigenous algorithm for extracting phonetically rich sentences, on-line and off-line recording facility, tacking code-mixed data, quality checking & control are unique features that enable the framework collecting data from remote and rural parts of the country. It also highlights issues and hurdles faced in collecting sample data and addressing them.

Keywords: Crowdsourcing, Under Resource language, Phonetically Rich sentences

1. Introduction

Current language and speech technologies are data driven. This is due to the fact that model built over data needs to collect evidences from different instances. Large amounts of annotated speech data are needed to model the effects of different sources of variability. An axiom of speech research is - there are no data like more data.

India is one of the largest and fastest growing markets for digital consumers, having 560 million internet subscribers in 2018 (TRAI, 2018), second only to China. According to McKinsey report 2019, India is one of the largest and fastest growing markets for digital consumers, and India's lower-income states are bridging the digital divide, and the country has the potential to be a truly connected nation by 2025. The experiment described in this paper presents exploiting the use of smartphones for collecting huge and varied speech data. The speech data collection has crossed the boundaries of studio based sequential recordings to crowd-source based parallel recordings. This paper details such a framework which can be used for multiple languages, utilizing common phone-set for Indian languages, indigenous algorithm for extracting phonetically rich sentences, on-line and off-line recording facility, tacking code-mixed data, quality checking & control while recording, and enabling collection of voice data from rural and remote areas at the user's choice of time in multiple sessions.

2. Related Works

Switchboard (Godfrey et al., 1992)[1], Fisher (Cieri et al., 2004)[2] and Broadcast News (Garofolo et al.,1997) corpora[3] are some of the prominent high-quality corpus. These all have taken a long time and huge effort and are considered landmark in speech technology field.

In Indian languages Speech Database for Hindi, Indian English and Bengali languages have been recorded by 1500 speakers in each language covering different environment conditions, Age Groups and gender distribution. The speech data is collected through IVR mechanism over mobile recorded speech data, sampled at 8 KHz. The crowd-sourcing mechanism has been is use for sometime Crowdee, CowdFlower etc[4,5]. The effort and framework presented here advocates and uses an indigenous corpus design methodology, yet provides the advantage of speedy and near to real life scenario data collection which is desired for building robust ASR system. The audio recordings are not limited to some specific microphones or recording devices. This also helps in collection of speech corpus covering a variety of recording devices and thus generalizes the speech corpus.

3. Architecture

The client application works on the user's Smartphone. The overall communication happens over the HTTP protocol. The Android based App 'मेरी भाषा मेरी वाणी' [6] (My Language My Voice) facilitates user to read out the sentence or phrase displayed over the screen. The session is maintained once a user is connected to server. On completion of recording, the user has the option to verify it through playing it before final submission. On pressing submission button, the preliminary level validation is carried out on the client side and on passing the preliminary check the recording is submitted to the server and next prompt is received. The App is Android based and works on majority of smartphones being used in India.

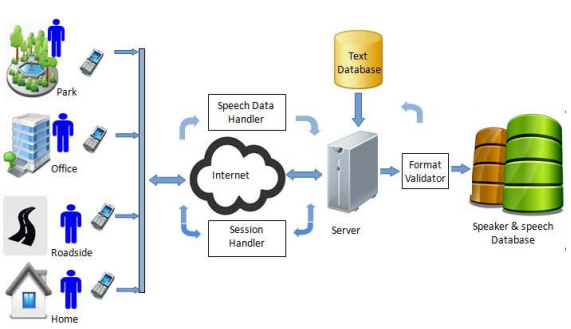


Fig. 1. Client-Server architecture for speech data collection

4. Text Corpus Development

Cleaning and filtering of corpus: The html boilerplate are removed from documents. The text is filtered on the basis of character encoding (UTF-8)/ASCII to adhere to the consistency of English alphabets and numerals. Additional portal specific contents for example URLs, emails, addresses etc. are dropped. Identification of improper syntax e.g. existence of invalid bigrams/character combinations has been done. Sentences with foreign word presence are filtered so as to have a good quality monolingual corpus. Inadequate sized sentences and words are identified and removed. Invalid Unicode patterns in sentences and Out of range character in sentences are not considered. Duplicate sentences along with duplicate punctuations are also removed.

An algorithm was developed to extract Phonetically/Grapheme Rich Sets from any corpus. The thresholds for the same was made configurable in order to limit sentences. The grapheme ratios were maintained to be a factor of original representative corpus. Maintaining such ratios provided a more natural way of increasing phonetic richness of corpus.

Buckets creation: In order to include the extracted phonetically rich sets a different algorithm was worked on

- The sentences were categorized into buckets based on sentence lengths
- The buckets were sorted according to their grapheme richness.
- Also, a separate data structure was maintained in order to a track each grapheme present in a sentence along with the counts.
- Once the data structure is ready, algorithm adaptively decides the rare graphemes, and starts to pick one sentence each bucket.
- Works on to extract sentences till a specified threshold of all graphemes is reached.
- After that, works only to enrich the set of graphemes that are rare.
- Thus, ensuring proper distribution of sentences across the prompt sheets, taking into consideration readability and grapheme richness.

5. Components of a general prompt sheet

The database contains sentences, guided prompts and unguided prompts/queries. The sentences are designed having minimum phonemic distribution and falling within certain length. Phonetically rich sentences of length less than 16 words [5] are collected using greedy algorithm and proof reading is done to ensure correctness of data. Some sentences were framed manually and are introduced to optimize the set. The Table 1 lists the items covered in the speech database.

Speech Database contents
Phonetically rich sentences collected from news papers, books, BTEC sentences and web
Proper Names - Indian males & females, Cities/States/Countries
Visiting Places, Monuments, Parks, famous buildings, Airport names, Airline names, Railway Stations names, Train names etc.
Unguided Queries with expected responses – - Isolated Digits - Connected Digits - Date & Time Vocabulary - Vocabulary - Currency and Money - Measurements - Yes/No utterances
Guided Prompts containing – - Isolated Digits - Connected Digits - Date & Time Vocabulary - Currency and Money - Measurements Yes/No utterances
Silence to capture background noise

Table. 1. Speech Database contents

The digits and numbers vocabulary covers telephone/mobile numbers, PIN codes, credit card numbers and natural numbers, date and time expressions contain - months, days, holidays, time, Proper names contain (person names and geographical entities like cities, states and countries). Guided prompts list out the entities in word format to get recordings in the planned way and unguided prompts give freedom to the speaker to speak in natural manner. Unguided prompts appear before the guided prompts to avoid biasing. For example, in normal scenario, the mobile numbers or monetary values may be spoken by a person in more than one way. In many real cases, it can be easily observed that people even do the code switching of the language also. In un-guided scenario, it has been tried out to have the queries, so that person answers them in his/her preferred/casual manner. While in guided scenario the speaker is provided the way in which he/she is supposed to speak.

6. Quality Control during recording

The client application also performs some of the quality checks while recording itself. These include presence of short silence zone before the start of a prompt and after the end of a spoken prompt which is to clipped utterance of prompts. The second quality check is SNR ratio, if it is not above a threshold the user is asked to either move to some lesser noisy place or speak a bit louder.

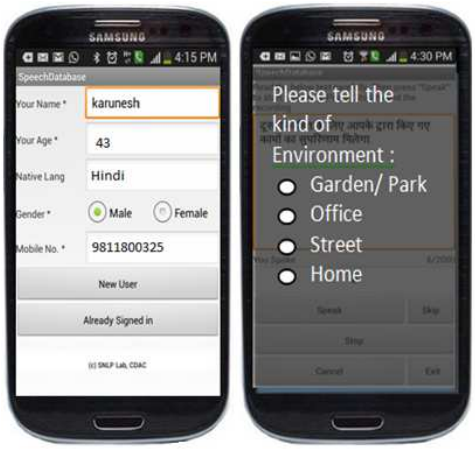
7. Interface for recording

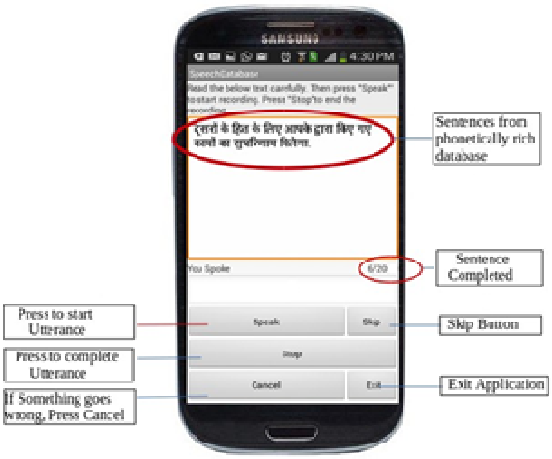
The speaker is provided with a user friendly interface on the mobile. The Registration phase captures the Meta information about the speaker, such as his demographic details, environmental details and equipment information.

7.1 Client Application

Client application is a mobile app facilitates user with an interface for recording voice as well as on screen display of sentences to utter. Before initiation of the recording session, the user has to register. As part of the registration, user provides Meta information like Name, Age, Gender, Mother Tongue, Native Place, Mobile Number and environment in which they are currently accessing the App (Fig. 2).

After Registration phase, the recording session starts. The sentences/prompts keep on appearing on the mobile screen, one after another, and user is also provided facility to skip the sentences/prompts he/she feels uncomfortable and can repeat recording if not satisfied. The speaker can record in single or multiple sessions as per his/her convenience (Fig. 3). The sentence/prompt wise speech data is transferred to the server, as soon as he/she goes for the next.





App has two modes of operations:

a. Online mode: In this mode, speaker's recording is sent to server upon completion of each utterance, and next sentence appears only when transaction is successful.

b. Offline mode: In this mode, where internet connection is not available, user is presented with pre-stored set of sentences in the App. He can record his utterances one by one and complete his session. Later on, whenever there is internet facility is available for him, the recordings can be uploaded in batch.

8. Recording Workflow

Here, we list out the steps being followed in Recording process.

8.1 Selection of speaker

Though the App is accessible to all, yet in initial phase, we have provided access to various institutions where a number of speakers can contribute their voice. For this institute level login facility has been extended for managed crowd-sourcing.

8.2 Registration and Meta-data information

A basic electronic record of the speaker's personal information is entered into the mobile device, including Age, Mother Tongue, Place and level of education, mobile number and data collection consent etc.

8.3 Training

The speaker is briefed and trained to utter different prompts after indication beeps etc. Sometimes, sample recordings are done. Videos / Audios of pre-recorded sessions are played. First two prompts are provided for the purpose of training and getting feel of the whole process.

8.4 Recording

Actual prompts recording takes place after successful training and confidence level of speaker.

8.5 Reward

Upon completion of the recording session, the primary validation of recorded voices is also carried out to ensure the proper recordings. The respondent is rewarded as per prior agreement.

9. Observations & Challenges

The most of the recordings were found by the speakers in the age-group of 18-30 years. Some of the issues observed are listed below:

a. Speaker sometimes moved mobiles away from mouth (to read the prompt from mobile screen) while recording, so intra-utterance variation in amplitude was observed.

b. However, not fumbled, but broken pronunciation of difficult words was observed like फास्फोलिपिड्स, कार्बोहाइड्रेड-उपापचय, डाइसल्फाइडेज.

c. Long silences in-between and at the end of sentences captured unnecessary noise.

d. Some speakers' recording were not in natural as expected, though this ratio is quite low.

10. Future Work

The paper presents here the use of crowd-sourcing through the most common device mobile for speech data collection in collaborative and cost effective manner. The speech data collection span got reduced. It may also help in maintaining naturalness in the speech, as people felt more comfortable speaking to the mobile. The system comes with easy to use interface and prompts the speaker with sentence/text to be read/spoken by him/her. The bar above the display window provides the instructions to guide the speaker. This helps in simultaneous recordings with minimum manual handholding. The speaker is able to complete his recording in split sessions. In speech database collection through this way, we do not have any control on microphone type; distance of phone, ambient noise etc., yet this comes with the advantage of gathering the speech data in close to natural way of speaking. The data collected in this way is representative of the actual test data which ASR would be subjected to in the web or mobile based application. As a future work this data would be used to improve the acoustic models of the existing ASR system [7]. The framework used in this experiment is configurable and independent of the language and will also be used for other languages.

11. Conclusion

It is time to save cultural heritage of language and associated culture in India. There are many languages which are almost zero resourced. New tools and technologies combined with innovative approach to attract Crowdsourcing can really help to revive many endangered languages. Through "मेरी भाषा मेरी वाणी", we have planned to create Speech resorces for Punjabi, Bengali, Assamese and Gujarati in the first phase. This languages are less-resourced to start with. Later on we will cover very less or zero resourced languages.

12. Acknowledgements

We thank the speakers who have contributed their speech data. Thanks are due to Mr. Dipankar Ganguly for developing algorithm for extracting phonetically rich sentences. We thank our Executive Director and management for providing conducive environment for performing this task.

13. Bibliographical References

1. Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). SWITCHBOARD: telephone speech corpus for research and development. In Proceedings of Acoustics, Speech, and Signal Processing (ICASSP-92). IEEE.
2. Cieri, C., Miller, D., and Walker, K. (2004). The Fisher Corpus: a resource for the next generations of speech-totext. In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC). European Language Resources Association.
3. Garofolo, J., Fiscus, J., and Fisher, W. (1997). Design and preparation of the 1996 Hub-4 Broadcast News benchmark test corpora. In Proceedings of the DARPA Speech Recognition Workshop
4. https://app.crowdee.de/
5. https://data.world/crowdflower
6. https://play.google.com/store/apps/details?id=in.cdac.crowdapp
7. Sakriani Sakti, Michael Paul, Andrew Finch, Shinsuke Sakai, Thang Tat Vu, Noriyuki Kimura, Chiori Hori, Eiichiro Sumita, Satoshi Nakamura, Jun Park, Chai Wutiwiwatchai, Bo Xu, Hammam Riza, Karunesh Arora, Chi Mai Luong, Haizhou Li, "A-STAR: Toward Translating Asian Spoken Languages", Computer Speech and Language, Special Issue on Speech-to-Speech Translation, Volume 27, Issue 2, pages 509-527, 2013.Indian telecom services performance indicator report, June-September 2018, Telecom Regulatory Authority of India.
8. Marge, M., S. Banerjee, A. I. Rudinicky, "Using the Amazon Mechanical Turk for Transcription of Spoken Language", *In Proc. of ICASSP*, 2010.
9. Yang, Z., B. Li, Y. Zhu, I. King, G. Levow, H.M. Meng, "Collection of user judgments on spoken dialog system with crowdsourcing, *In Proc. of SLT*, 2010.
10. Lane, I., M. Eck, K. Rottmann, A. Waibel, "Tools for Collecting Speech Corpora via Mechanical-Turk", *In Proc. of Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010.
11. Arora Karunesh, Arora Sunita, Agrawal S. S., Paulsson Niklas, Choukri Khalid. "Experiences in Development of Hindi Speech Corpora based on ELDA standards*". In Proc. of the Oriental COCOSDA,* 2006.
12. R. Molapo, E. Baranard, and F. de Wet, "Speech data collection in an under-resourced language within a multilingual context," in 4th International Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU), 2014.
13. L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," Speech Commun., vol. 56, pp. 85–100, Jan. 2014

.

Proceedings, pages 87–90
Paris, UNESCO Headquarters, 5-6 December, 2019.

IARPA's Contribution to Human Language Technology Development for Low Resource Languages

Carl Rubino
IARPA
4600 Sangamore Road, Bethesda MD 20816 USA
Carl.Rubino@iarpa.gov

Abstract

With the goal of advancing the state of natural language processing development in constrained resource conditions, using machine learning and repeatable methodologies, the Intelligence Advanced Research Projects Activity (IARPA) launched two multi-year research programs: Babel and MATERIAL. This paper details the philosophy and objective behind each program to advance technologies in this area and introduces the program corpora available to stimulate related research.

Keywords: Low Resource Languages, Cross-language Information Retrieval, Machine Translation, ASR, corpora, Cross-language summarization, machine learning, speech technology, AQWV

Résumé

Iti gandat ti panagpadur-as iti agdama a kasasaad ti automatiko a panagproseso ken panagsursuro kadagiti pagsasao no agkurang ti datos a maikurri, insayangkat ti IARPA ti dua a dakkel a programa ti panagsukisok a nagpaut iti sumagmamano a tawen: Babel ken MATERIAL. Iburayko ditoy ti panggep ken pilosopia dagiti dua a programa iti panangparang-ay kadagiti kastoy a teknolohia. Ilawlawagko met ti amin a korpora dagiti programa nga imbunong ti IARPA tapno makadur-as pay ti panagsukisok iti daytoy a tay-ak.

1. Introduction

Advances in natural language processing (NLP) have continued to progress substantially since the advent and application of deep learning methods. Several applications in this realm have benefited from breakthroughs resulting from sustained refinements and methodological evolution. However, despite impressive progress in performance for multiple human language technology areas, the gap in performance between English and other languages suggests that the application of these novel techniques to new languages does not necessarily portend success. Deep learning methods consistently require much more data than is usually available for the majority of the world's languages. Moreover, deep learning methods are often impacted by noise in training data more than traditional machine learning methods. Recognizing the need to improve human language technologies for lower resource languages, the Intelligence Advanced Research Projects Activity (IARPA) invested in multiple relevant research endeavors.

To address unsolved problems deemed important to the Intelligence Community, IARPA, the research wing of the U. S. Office of the Director of National Intelligence, employs a competitive bid process to mobilize the best talent worldwide to work on our research programs. That competition process is our solicitation of proposals in response to a Broad Agency Announcement (BAA) of our research programs. Typically, we fund multidisciplinary teams to address the research comprehensively. Selected teams work collaboratively and competitively to tackle the challenges and advance our understanding of the problems, frequently with cross-disciplinary solutions. To propel research toward solutions that work, IARPA measures Performer Team progress via a rigorous and well-defined metric-based evaluation that is often a product of both initial foresight and lessons learned from active engagement in the research activities.

In the realm of human language technology, IARPA has launched a variety of research initiatives ranging from small-term studies and seedlings (research efforts of less than a year) to five multi-year research programs. Of the five programs, three are complete: SCIL (Socio-cultural Content in Language for social role and goal discovery), METAPHOR (analysis of metaphor to gain insight into interpreting cultural norms), Babel (Speech Recognition); and two are currently ongoing: MATERIAL (Machine Translation and Cross-language information Retrieval) and BETTER (Cross-language Information Extraction and Retrieval). We will only cover MATERIAL and Babel here, as they involve research with low resource languages in low resource conditions. These endeavors greatly expanded the IARPA portfolio of complex, multidisciplinary programs[1], but most importantly for the NLP community, pioneered a new evaluation paradigm to measure NLP progress, and provided several large annotated datasets accessible to the community to encourage continued research in this area.

[1] IARPA has a diverse research portfolio encompassing research from a variety of disciplines, including math, physics, linguistics, biology, neuroscience, political science, and cognitive psychology.

2. The Babel Program

The Babel program, the brainchild of Dr. Mary P. Harper, was launched in 2011, to address two technological gaps in speech technology (Harper 2011). At the program's inception, mature technologies for low resource languages to process speech in a meaningful way for keyword search (KWS) was non-existent, and the time and resources required for system development to address this problem were beyond the reach of most research institutions. To address these shortcomings, IARPA contracted multiple multinational "performing teams" of experts who competed to develop agile and robust methods for supporting effective keyword search over massive amounts of recorded- speech in foreign languages.

	BP	OP1	OP2	OP3
Practice Languages	**4 languages:** Cantonese, Pashto, Tagalog, Turkish	**5 languages:** Assamese, Bengali, Haitian Creole, Lao, Zulu	**6 languages:** Cebuano, Kazakh, Kurdish, Lithuanian, Telugu, Tok Pisin	**7 languages:** Amharic, Dholuo, Guarani, Igbo, Javanese, Mongolian, and Pashto (revisited)
Surprise Language	Vietnamese in **4 weeks**	Tamil in **3 weeks**	Swahili in **2 weeks**	Georgian in **1 week**
Training Data Limits	80 hours, 10 hours (with dictionary)	**60 hours**, 10 hours (with dictionary)	**40 hours, 3 hours, Select 3 hours (no dictionary)**	40 hours (no dictionary)
Recordings	Mixed Environment Telephone	Mixed Environment Telephone **& Microphone (2 types)**	Mixed Environment Telephone **& Microphone (2 types)**	Mixed Environment Telephone **& Microphone (7 types)**
ATWV Target	0.3 or greater	0.3 or greater	0.3 or greater	**0.6 or greater**
WER Target	N/A	N/A	N/A	**50% or less**
Meet Target with:	80 hours	**60 hours, 10 hours**	**40 hours, 3 hours, Selected 3 hours**	40 hours
What was in the BAA?	80 hours (with dictionary)	60 hours (with dictionary)	40 hours (with dictionary)	40 hours (with dictionary)

Figure 1 Babel languages and targets per period

As a way to ensure efficacy of the search tools, and portability of the methods used to build them, IARPA used diverse languages from multiple language families, and real-world recording conditions in the training and evaluation. Data were collected and consistently transcribed in-country using normalized conventions. Extensive vetting of the resources was performed by the University of Maryland's Center for Advanced Study of Languages (CASL), with particular attention to languages with less standardized orthographies. IARPA contracted MIT Lincoln Labs as a Test and Evaluation partner to process the corpora for effective evaluation, and develop speech recognition reference systems to baseline pre-program state-of-the-art performance to set meaningful metric thresholds for the performing teams. The National Institute of Standards and Technology (NIST) was tasked with developing the program evaluation metric and with conducting regular evaluations of the performing systems.

Automatic Speech Recognition (ASR) training data created for the program included 40-80 hours of transcribed speech in each language, provided to program participants upon the kickoff of each new language to jump start their research. Datasets used in Babel consisted of natural conversations of at least two thousand speakers, recorded under various microphone and service provider conditions (See Figure 1).

The metric used to evaluate progress towards the program goal was ATWV (Actual Term Weighted Value), a detection metric that, for each foreign keyword query, awards true hits and penalizes false alarms and misses, with relative weights set by IARPA, then averages the scores over an entire query set[2]. This detection metric was deemed most suitable to measure performance as it ensured each query would have equal weight regardless of its relevance probability. System developers would choose a decision threshold of their probability computations for their "Actual" decisions to maximize this metric, and IARPA would use these scores to determine progress and the status of each team in the program. In the fourth and final program period, KWS systems were also expected to achieve a word error rate (WER) of fifty percent or less.

As Babel progressed, teams faced increasingly more challenging program goals in shorter lengths of time and with less training data, leveraging approaches they learned to mitigate issues associated with resource and transcription dearth. Multilingual features and acoustic models proved effective, as well as grapheme based acoustic modeling and cascaded adaptor grammars to better resolve new words not seen in the training data.

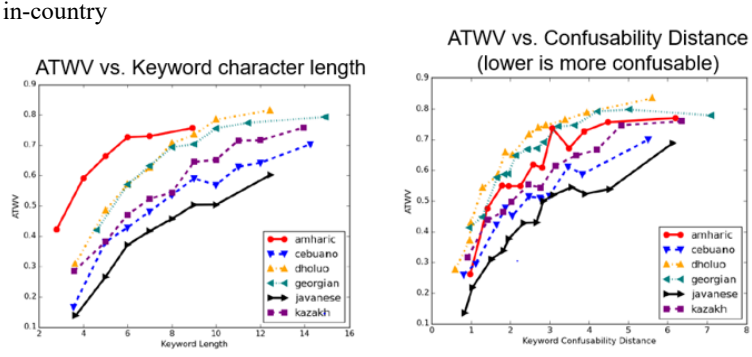


Figure 2 ATWV Correlations as reported by BBN

IARPA and Performing Teams investigated multiple issues to understand differential progress between systems, configurations and languages (Hartmann et al 2017). A number of variables seemed to correlate with overall system performance. Variables investigated included word error rate performance, average word length, keyword

[2] OpenKWS Evaluation Plan, https://bit.ly/36z6BsR

confusability distance (weighted keyword length where lower values are more confusable, See Figure 2), graphemic error rate calculated against the training data (Figure 3), speaking rate (phones per second), keyword frequency, and inter-annotator agreement in the transcriptions. Measuring these details was not just critical to understanding of performance degradations to allow research teams to appropriately compensate for them, but also to inform the program manager of optimal language choices for subsequent periods of the program based on predicting expected performance, as well as query construct design.

To optimize the methods used to both advance the science and evaluate performance, a good program design was critical. A strategic and meaningful choice of languages proved to be most important. A number of criteria were considered prior to language selection, ranging from language-specific features (phoneme inventory, morphological complexity, orthography, syntax, dialectal variation, typological uniqueness, and genetic relationships) to accessibility and the cost of the collection. For the latter measure, Dr. Harper as the Program Manager took into account the number of native speakers, quality of phone connections, availability of linguistic expertise for transcription, and the political stability of the region to assure a safe in-country collection. In addition, it was strongly desired that the languages released in Babel were not spoken by members of the Performer Teams.

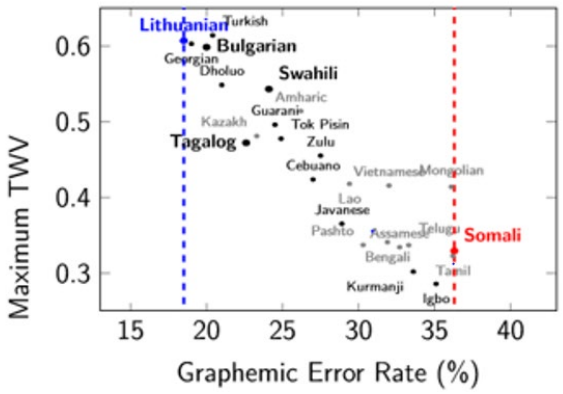


Figure 3 GER was a strong predictor of performance

Prior to each data collection, a "Language Specific Peculiarities" (LSP) document describing unique and essential properties of each language was produced to ensure consistency in translation, a well-balanced collection in terms of dialect coverage, and to optimize keyword selection for the final evaluation. Teams were also provided with pronunciation lexicons for words in the training set. These lexicons were converted to a finite state transducer (FST) so each orthographic word would be mapped onto a set of reasonable pronunciations. The Test and Evaluation team also created confusion matrices for words whose pronunciations were similar but not identical. The LSP documents, pronunciation lexicons, and training data for the twenty six languages released in this program are available from IARPA, and via the University of Pennsylvania Linguistic Data Consortium, http://-ldc.upenn.edu

3. The MATERIAL Program

Leveraging lessons learned from Babel, the MATERIAL (Machine Translation for English Retrieval of Information in Any Language) program was launched in 2017 to evaluate performance on a much wider array of human language technologies to include cross-language information retrieval and summarization (Rubino 2017). Performers on this program built systems that can retrieve foreign language speech and text documents responsive to domain-constrained English queries, and provide evidence or relevance, in English, to both the query string and its domain. This novel evaluation measured performance, not on each underlying technology involved, but on two functional and unified End-to-End capabilities in a way these technologies were had not been evaluated before. The first performance score, AQWV (Averaged Query Weighted Value), measured the effectiveness of cross-language information retrieval systems. The second, measured the End-to-End performance (E2E AQWV) of systems as judged by humans on their cross-language summarization capability with a novel crowd-sourced evaluation methodology utilizing the Amazon Mechanical Turk platform designed and executed by Tarragon Consulting[3]. CLIR AQWV scores could be improved in the E2E evaluation if summarization systems provided enough evidence for the human judges to reject false alarms. However, the CLIR AQWV scores could also degrade if the summaries were not of sufficient quality to convince the judges to retain true positive documents.

To effect this evaluation paradigm, unique datasets were compiled to include domain-annotated documents in six genres, queries of various types designed to probe various analytic dimensions, and relevance decisions for each query against the program documents. The query typology developed for MATERIAL allowed IARPA to probe each system's ability to handle ontological concepts, and to resolve ambiguity resulting from polysemy, homophony, and/or named entities.

As of December 2019, five languages were released for evaluation. For the Phase I period languages (Swahili, Tagalog and Somali), systems were also required to identify eight domains: Government and Politics, Lifestyle, Business and Commerce, Law and Order, Physical and Mental Health, Military, Sports, and Religion. For Phase II, three additional languages will be evaluated: Lithuanian, Bulgarian, and a surprise language to be released in January 2020. IARPA plans to release 3 more languages in the final Phase III of the program.

Like the Babel program, language expertise procurement was disallowed to drive approaches that leverage machine learning. Likewise, each period of the program had a development stage in which performing teams worked with multiple "practice languages" to develop their methods. The development stage was followed by an evaluation stage, in which progress was officially evaluated on a

[3] MATERIAL Evaluation Plan, https://bit.ly/2oRT923

"surprise language". The evaluation dataset was partitioned into three temporal epochs, corresponding to different query sets and new domain releases. Cross-language query-biased summaries, providing relevance justification for each retrieved document, were evaluated at the end of each phase by English-speaking crowd sourced judges. This fundamental development and evaluation cycle is illustrated in Figure 4. For Phase I of the program, IARPA released Swahili and Tagalog as the practice languages. The surprise language was Somali.

MATERIAL documents were collected in two modes (text and speech), and six genres (news text, monologic social media text, and topical text, conversational speech, broadcast news audio, and topical audio). A challenging aspect of the evaluation was the planned mismatch condition between the training and evaluation conditions.

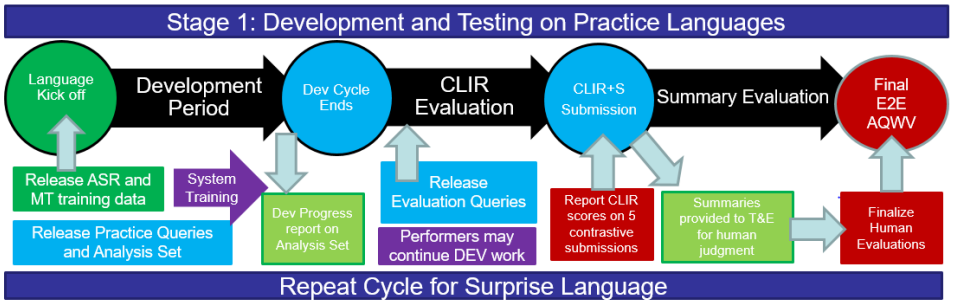


Figure 4 Development and Evaluation Cycle for each MATERIAL Language

At the release of each new language, IARPA distributed "Training Build Packs". These consisted of fifty hours of transcribed telephony for ASR development, phonetic lexica, a parallel corpus for MT training consisting of approximately 800K English words from news texts translated and aligned at the sentence level, and a language description that detailed language-specific peculiarities pertaining to both the text and audio collection, transcription, and the language itself. To challenge teams to create portable solutions, not all speech and text genres in the evaluation condition were represented in the training data. Teams developed novel methods to cope with the mismatch. Considering the dearth of training provided for each language kickoff, teams were expected to complement the IARPA-provided build packs with their own data harvests.

Beyond the training datasets, the remaining documents were partitioned by NIST and MIT Lincoln Laboratories into three other sets, optimized to achieve a reasonable training/evaluation balance. A "devtest" set of documents with relevance annotations for each query was provided to each team to allow them to locally score their systems on CLIR via the AQWV metric. Teams were discouraged from using this set for training to better understand the results of component engine work and joint process optimization. An analysis set of documents was provided with the devtest and after each evaluation for glass box testing to analyze errors. The analysis set was fully translated; audio documents in this set were transcribed with the conventions employed in the training build pack. Teams were allowed to closely scrutinize this set to help diagnose ASR, MT and CLIR errors and understand progress. Finally, a blind evaluation set was divided temporally into three epochs, and scored at NIST to reveal final results.

Although language-independent machine learning approaches are desirable to effect quick capability ramp up, it was obvious that not all languages are created equal. Different strategies were employed per language to optimize AQWV. Each system analysis yielded interesting observations. We found that vowel removal strategies that helped Tagalog, hurt Swahili and Somali. Translation lattices that helped Somali and Tagalog did little to improve results on Somali. Stemming during retrieval and translation benefited Swahili and Somali but not Tagalog. IARPA also investigated dataset peculiarities per language. Vocabulary growth, OOV (out of vocabulary) words in evaluation not in training, sentence perplexity and audio clipping were measured as factors that may affect performance. It was immediately evident that the high vocabulary growth rate in Lithuanian did not result in lower AQWV scores. The relative wealth of resources available for this language compensated for this perceived difficulty of the expanding vocabulary. The next language to be released will offer additional challenges to include rampant orthographic and dialectal variation.

4. Conclusion

With proper design resulting from well-informed planning and sustained collaboration with expert scientific teams, the U.S. Government can launch programs to tackle seemingly impossible challenges in NLP. Teams in the Babel program proved that an effective keyword search capability can indeed be developed in one week. MATERIAL teams are exploring novel methods to enable cross-lingual semantic search of text and audio outside the traditional realm of machine translation under realistic constraints that mirror the current problem space. It is our goal that the insights and lessons we have learned from our investments and work are applied by the community as we propel the research to take on more ambitious problems in the future. IARPA invites the research community to learn from both our progress and mistakes, and to profit from the datasets we will disseminate to see how much further they can go. IARPA also welcomes the community to propose high-risk high-reward ideas for new research in the NLP domain.

5. Acknowledgements

My thanks to Ilya Zavorin and Catherine Cotell for their helpful comments on a previous draft of this paper.

6. Bibliographical References

Harper, M. (2011). Babel BAA. https://bit.ly/2KvdDoY.

Hartmann, W, D. Karakos, R. Hsiao, L. Zhang, T, Alum. and Gertz, M. Alumäe, S. Tsakalidis and R. Schwartz (2017). Analysis of Keyword Spotting Performance Across IARPA Babel Languages, ICASSP'17, pages 5765-5769.

Rubino, C. (2017). Material BAA. https://bit.ly/37gKhV9.

Proceedings, pages 91–94
Paris, UNESCO Headquarters, 5-6 December, 2019.

Google Crowdsourced Speech Corpora and Related Open-Source Resources for Low-Resource Languages and Dialects: An Overview

Alena Butryna, Shan-Hui Cathy Chu, Işın Demirşahin, Alexander Gutkin, Linne Ha[†], Fei He, Martin Jansche[†], Cibu Johny, Anna Katanova, Oddur Kjartansson, Chenfang Li[†], Tatiana Merkulova, Yin May Oo[†], Knot Pipatsrisawat, Clara Rivera, Supheakmungkol Sarin, Pasindu de Silva, Keshan Sodimana[†], Richard Sproat, Theeraphol Wattanavekin, Jaka Aris Eko Wibawa[†]
Google Research, Singapore, United States and United Kingdom
{alenab,oddur,rivera,mungkol}@google.com

Abstract

This paper presents an overview of a program designed to address the growing need for developing freely available speech resources for under-represented languages. At present we have released 38 datasets for building text-to-speech and automatic speech recognition applications for languages and dialects of South and Southeast Asia, Africa, Europe and South America. The paper describes the methodology used for developing such corpora and presents some of our findings that could benefit under-represented language communities.

1. Introduction

Historically speech and language technology research has focused on a few major Indo-European languages, along with Mandarin and Japanese. The past decade, however, has seen an increased focus by the speech and language research community and technological companies on addressing the plight of low resource and especially the endangered languages. According to various sources, such as Ethnologue (2019), roughly 40% of 5,000 to 7,000 languages spoken today are classified as endangered. The shift of focus is partly due to the awareness of the importance of preserving and documenting the languages which are at risk of losing its last native speakers due to the shift to other dominant languages or disappearance of the communities altogether. In addition to the critically endangered languages, there are hundreds of languages with large native speaker populations which are classified as low-resource (sometimes low-density) due to the lack of linguistic resources necessary for advancing research and technological innovation. Furthermore, the Internet is creating a large and growing divide between languages that are represented in technology and those that are not: it is estimated that only 5% of the world's languages are accessible on the Internet.

In this paper we offer a brief overview of a linguistic program which aims to provide free speech resources in regions with fast growing Internet communities but few publicly available linguistic resources. So far the efforts have mostly focused on statutory national or provincial languages, with the overall goal of investigating the methods to scale our approach to many smaller regional languages in the locales of interest. One of the primary goals of the program is to develop an accessible and replicable methodology that any local community of technologists can use with available open-source solutions to build custom applications utilizing our released resources or to construct their own resources for a new language. At the same time, it is important to make sure that the quality of the resulting solutions built using this methodology are on par with the systems for better-resourced languages.

Another component of this program deals with the construction of corpora for low-resource dialects of well-resourced languages. More often than not, the assumption exists that a local community is adequately served by providing the speech technology built using the dominant dialect, despite the dialects significantly diverging to the point of being mutually unintelligible (e.g., High German vs. Swiss German). An interesting part of this process is investigating the optimal strategies for constructing local dialect-specific resources that build upon the existing well-resourced language resources in a way that adequately serves the local community.

The program focuses on developing the resources for two types of applications: automatic speech recognition (ASR) and text-to-speech (TTS), both of which are crucial components of modern technological ecosystems for any given language. These applications have different resource requirements: the modern ASR systems typically require more data from as many speakers as possible, while the TTS systems ideally need higher-quality recordings from fewer speakers but with well articulated speech. Additional resources, such as text normalization grammars for converting non-standard word tokens to natural language words and models of phonology are often needed as well. These are highly language-specific and require considerable linguistic expertise to develop.

2. Program Overview

2.1. Selection of Languages and Dialects

The languages and dialects selected for the program are broadly based on two selection criteria. The first goal is to increase the availability of open-source speech resources in the regions which were identified as important (in terms of number of speakers, Internet penetration and cultural significance for the region) and yet considered low-resource. The initiatives for constructing linguistic resources for such languages, whether from local or foreign governments or

[†]The author contributed to this work while at Google.

big technological companies, can effectively tip the balance and cause the language to become well-resourced, as happened with Modern Standard Arabic in the course of the last twenty years. The second criterion involves selection of languages from diverse language families so that the generality of resource collection paradigm can be optimized based on the exposure to different linguistic and operational requirements and the findings shared with the community. The TTS and ASR corpora collected so far consist of over 1,500 hours of speech and are freely available online hosted by Open Speech and Language Resources (2019) under unencumbered license (Creative Commons, 2019).

South Asia The South Asian languages selected for the program include the languages from Indo-Aryan and Dravidian language families. The Indo-Aryan languages selected so far include two dialects of Bengali (India and Bangladesh), Gujarati (India), Marathi (India), Nepali (Nepal) and Sinhala (Sri Lanka). The set of Dravidian languages included by the program include Kannada (India), Malayalam (India), Tamil (India) and Telugu (India). According to various estimates, these languages have a combined population of about 706 million native and second-language speakers. From a research standpoint, these languages are very interesting to work with: they exhibit considerable variation within each language family, but at the same time also have considerable similarities across both language families.

Southeast Asia The set of Southeast Asian languages we selected includes Burmese (Myanmar) from Sino-Tibetan language family, Khmer (Cambodia) from Austroasiatic language family, and Javanese (Indonesia) and Sundanese (Indonesia) from Malayo-Polynesian language family. These languages are natively spoken by about 178 million people across the region. The language families in this set are very diverse and yet exhibit considerable influence from their neighbors from other language families in both South and Southeast Asia.

Africa Four out of eight official languages of South Africa were selected: Sesotho, Setswana, Xhosa (from a Bantu language family) and Afrikaans (Indo-European). These languages have a combined speaker population of native and second-language speakers of about 62.5 million. In addition we selected Nigerian English as one of the largest and yet low-resource dialects of English on the continent.

Europe and South America To increase the coverage of Indo-European language family among the selected languages, a set of three regional languages of Spain with combined population of about 14 million speakers were selected: Galician and Catalan, both of which belong to Ibero-Romance group, and Basque, which is a language isolate. As part of our work towards improving the availability of open-source speech resources for low-resource dialects and regional accents of the better served languages, we also collected speech corpora for six Latin American Spanish dialects (Argentinian, Colombian, Chilean, Peruvian, Puerto Rican and Venezuelan) and various dialects and accents of Irish and British English (Welsh English, Southern English, Midlands English, Northern English and Scottish English).

2.2. Methodology

Local Community and University Outreach It goes without saying that any collaborative corpora collection is greatly helped by the enthusiastic community of native speakers. Throughout the program we tried to enlist the help from local universities, technology and language enthusiasts wherever possible. This approach is illustrated by the data collection process for Javanese and Sundanese for which the collaboration with two local universities was established. For Javanese we worked with the Faculty of Computer Science at Universitas Gadjah Mada (UGM) in Yogyakarta, while for Sundanese a collaboration with the Faculty of Language and Literature at Universitas Pendidikan Indonesia (UPI) in Bandung was established. The universities assisted us with finding volunteers to help manage the data collection, as well as with the adequate recording environments. The university staff put us in contact with the student organizations which helped to disseminate the information about corpus collection and call for volunteers. A portion of the recordings was done at the student-run annual Computer Science exhibition event organized by students from the Faculty of Computer Science at Universitas Indonesia (Wibawa et al., 2018). A similar approach was followed in South Africa where we established the collaboration with Multilingual Speech Technologies group from the North-West University to assemble the speech corpora for four South African languages (van Niekerk et al., 2017). In other countries, such as Bangladesh, Cambodia, Myanmar, Nepal and Sri Lanka, we followed the same blueprint involving the participants from local universities in the processes of data collection and curation (Kjartansson et al., 2018). In Latin America, we worked with local Google Developer group representatives.

Software and Hardware Equipment For ASR data collections, we required recording applications capable of running on low-end smartphone devices. While initially we relied on a proprietary application, we later teamed up with University of Reykjavik in Iceland and migrated our ASR collections to use their open-source software (Petursson et al., 2016). Since TTS corpora requires higher speech quality, we went through several careful iterations to settle on a hardware combination that was lightweight and portable while providing the best possible quality for our purposes. Additionally, we made sure that the equipment in question was affordable to local communities. One of the configurations that was found to work well for us and that we recommend to others includes an ASUS Zenbook UX305CA fanless laptop, Neumann KM 184 microphone, a Blue Icicle XLR-USB A/D converter and a portable acoustic booth. The overall cost of this configuration, especially when reused for multiple data collections, is well below the cost of renting a professional recording studio.

Development of Recording Materials Open sourcing low-resource language speech corpora was of high priority since the inception of the program. Therefore, during the recording script development we made sure we use publicly available sources. For both ASR and TTS corpora Wikipedia text was used in the form of short sentences extracted at random (when available in the language of interest, otherwise crowdsourced translations thereof). In addition,

for TTS recording scripts further combination of three types of materials was prepared: (1) Handcrafted text to ensure a broad phonetic coverage of the language, filling in any gaps from Wikipedia, (2) template sentences including common named entities and numeric expressions in each language. These were obtained in collaboration with communities and partners and include celebrity names, geographical names, telephone numbers, time expressions and so on, (3) domain specific real-world sentences that could be used in product applications, usually covering navigation, sports or weather.

Recording and Quality Control Procedures The ASR corpora are collected using standard consumer smartphones. No specialized or additional hardware is used for the data collection. The collections are overseen by the volunteer field workers who are trained to guide the volunteer speakers. During the recording the audio is first saved to local storage on the device and then uploaded to a server once a connection to the Internet is established. This feature of the recording software is important because limited Internet connectivity potentially poses serious operational problems to the field workers, especially in remote areas.

Recording of TTS corpora poses different challenges because the goal is to collect high-quality and well-articulated speech samples in a portable recording studio. A short introduction is given to participants so that they can confidently operate the software during the recording session, alongside with the guidelines on the relative position of the volunteer speaker to the microphone to ensure consistency between different recording sessions.

Since none of the speakers recorded for TTS are professional voice talents, their recordings often contain problematic artifacts such as unexpected pauses, spurious sounds (like coughing) and breathy speech. All recordings go through a quality control process performed by the trained native speakers to ensure that each utterance matches the corresponding transcript, has consistent volume, is noise-free and consists of fluent speech without unnatural pauses.

Other Types of Linguistic Resources Building competitive ASR and TTS applications for low-resource languages typically requires the development of further linguistic resources in addition to corpora and our program takes this requirement into account. The necessary components required for building robust speech ecosystem for any given language typically include carefully designed phonological representation upon which the pronunciation lexicons can be based, the algorithms for generating pronunciations for words missing from the lexicon and a system for converting between non-standard word (NSW) tokens, such as numbers, and the corresponding natural language words. The development of these components typically requires native knowledge of the language and considerable linguistic expertise. Therefore we are making sure that any additional linguistic artifacts developed by the program are well documented and freely accessible (Google, 2016). Examples of such artifacts include phonological representation for Lao, pronunciation guidelines for Burmese and text normalization grammars for languages of South and Southeast Asia (Sodimana et al., 2018).

2.3. Emerging Lessons

Operational Lessons Throughout the program's lifetime we continuously discovered the positive impact the collaboration with local communities had on our data collection projects. This is partly due to local technologists, academics and open-source enthusiasts who understand well how the availability of technology in their local language can positively impact the life of a local community and often enthusiastically endorse initiatives such as ours. In our particular case, setting up the crowd-sourcing mechanisms locally would have not been possible without their support.

Furthermore, such collaborations often resulted in important contributions to other Google programs. For example, simultaneously with collecting the speech corpora in Bangladesh, Cambodia, Nepal, Myanmar and Sri Lanka, we hosted a series of media workshops, train-the-trainer sessions and translate-a-thons with universities and community groups, that aimed to educate people about how they can use the new Google Translate Community tool (Google, 2017) to improve the accuracy, understanding and representation of their language on the web. We also actively participated and contributed to various local conferences primarily focused around technology and the web.

Finally, we discovered that using moderately priced recording equipment was enough to collect the corpora of sufficient quality to suit the needs of local community and, at the same time, also be used in real Google products.

Research and Development Findings One of the first findings of this program is that the crowd-sourced TTS corpora works adequately in a single multi-speaker model (Gutkin et al., 2016). While the quality of the resulting model was somewhat below the quality of state-of-the-art commercial systems, at later stages of the program we discovered that the quality of such models can be significantly improved by combining multi-speaker crowdsourced corpora from multiple languages. This finding capitalizes on the notion of "linguistic area" by an eminent American linguist Emeneau (1956), where he defines it (p. 16, fn. 28) as "an area which includes languages belonging to more than one language family but showing traits in common which are found to belong to the other members of (at least) one of the families". Based on this observation we successfully built multilingual system based on the combined corpus of South Asian Indo-Aryan and Dravidian languages (Demirsahin et al., 2018) and Malayo-Polynesian multilingual system combining our Javanese and Sundanese corpora with a proprietary corpus of Indonesian (Wibawa et al., 2018). Furthermore, we found that our South Asian multilingual model was good enough for synthesizing the languages for which we had no training data, such as Odia and Punjabi.

Another lesson that emerged from applying the collected speech corpora in practical applications is the importance of freely available typological resources, such as PHOIBLE (Moran et al., 2014). During work on low-resource language technology, more often than not, the required linguistic (and in particular, phonological) expertise is hard to find, even among the native speakers. The availability of typological resources as a reference have significantly boosted our research and development efforts.

Important Challenges While working on this program, we have identified several important areas which need to be addressed in order to scale this work to many more languages and dialects which currently possess even fewer linguistic resources than the languages we have dealt with so far.

When it comes to developing speech corpora and applications, more often than not, the "one size fits all" approach does not work because different language families present very different challenges. Once the speech corpora are collected, the types of language-specific challenges that may block application development include the lack of large amounts of labeled training data for training word segmentation algorithms (e.g., Burmese, Khmer and Lao), lack of morphosyntactic taggers for smaller Slavic languages (e.g. Rusyn) required for proper function of text normalization and so on. Streamlining this process is highly non-trivial because currently no universal recipe exists.

Moreover, as we mentioned previously, development of linguistic resources requires considerable linguistic expertise, which is often hard to find. The deep learning end-to-end approaches, which have recently gained popularity (Toshniwal et al., 2018), offer potential workaround. Such systems can be adapted to smaller languages using transfer learning techniques (Chen et al., 2019). At the same time, such systems are notoriously data hungry and further techniques utilizing more data, including lower-quality data found online (Cooper, 2019), may be required. Furthermore, some form of linguistic knowledge is desirable in such systems due to occasional unpredictable errors they are prone to (Zhang et al., 2019).

3. Concluding Remarks

We have presented an overview of the program that helped collect and release 38 datasets for building TTS and ASR applications for languages and dialects of South and Southeast Asia, Africa, Europe and South America. The corpora collected so far consist of over 1,500 hours of speech and are freely available online. Partnering with local universities and communities in the region was crucial to the success of the program as it connected us with a lot of enthusiastic local contributors which in its turn resulted in collecting high quality data. We do hope that the described methodology and the released datasets will be utilized by the local communities to develop custom applications or to collect new datasets going forward.

There are still many endangered and low resource languages that we want to focus on in our program. Even though the program already allows to collect data for language resources development efficiently from an operational perspective, there are still challenges that need to be addressed at the development of linguistic resources stage so that the work can continue at scale. As of now, the program established a good foundation, however, there is still work to be done. We hope that these efforts will facilitate future research by the broader scientific community and will encourage others to apply our program methodologies and findings to benefit under-represented language research.

4. Bibliographical References

Chen, Y.-J., Tu, T., Yeh, C.-c., and Lee, H.-Y. (2019). End-to-end Text-to-speech for Low-resource Languages by Cross-Lingual Transfer Learning. *Proc. Interspeech 2019*, pages 2075–2079.

Cooper, E. L. (2019). *Text-to-Speech Synthesis Using Found Data for Low-Resource Languages*. Ph.D. thesis, Columbia University, New York.

Creative Commons. (2019). Attribution-ShareAlike 4.0 International (CC BY-SA 4.0). `http://creativecommons.org/licenses/by-sa/4.0/deed.en`.

Demirsahin, I., Jansche, M., and Gutkin, A. (2018). A Unified Phonological Representation of South Asian Languages for Multilingual Text-to-Speech. In *Proc. SLTU*, pages 80–84, Gurugram, India.

Emeneau, M. (1956). India as a Linguistic Area. *Language*, 32(1):3–16.

Ethnologue. (2019). Ethnologue, SIL International. `https://www.ethnologue.com`. Accessed: 2019-03-25.

Google. (2016). Google Internationalization Language Resources. `https://github.com/google/language-resources`.

Google. (2017). Google Translate Community. `https://translate.google.com/community`.

Gutkin, A., Ha, L., Jansche, M., Pipatsrisawat, K., and Sproat, R. (2016). TTS for Low Resource Languages: A Bangla Synthesizer. In *Proc. LREC*, pages 2005–2010, Portorož, Slovenia, May.

Kjartansson, O., Sarin, S., Pipatsrisawat, K., Jansche, M., and Ha, L. (2018). Crowd-Sourced Speech Corpora for Javanese, Sundanese, Sinhala, Nepali, and Bangladeshi Bengali. In *Proc. SLTU*, pages 52–55, Gurugram, India.

Moran, S., McCloy, D., and Wright, R. (2014). *PHOIBLE Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig. `http://phoible.org/`.

Open Speech and Language Resources. (2019). Open SLR. `http://www.openslr.org/resources.php`. Accessed: 2019-11-20.

Petursson, M., Klüpfel, S., and Gudnason, J. (2016). Eyra-speech data acquisition system for many languages. *Procedia Computer Science*, 81:53–60.

Sodimana, K., Silva, P. D., Sproat, R., Theeraphol, A., Li, C. F., Gutkin, A., Sarin, S., and Pipatsrisawat, K. (2018). Text Normalization for Bangla, Khmer, Nepali, Javanese, Sinhala, and Sundanese TTS Systems. In *Proc. SLTU*, pages 147–151, Gurugram, India.

Toshniwal, S., Sainath, T. N., Weiss, R. J., Li, B., Moreno, P., Weinstein, E., and Rao, K. (2018). Multilingual speech recognition with a single end-to-end model. In *Proc. ICASSP*, pages 4904–4908, Calgary, Canada. IEEE.

van Niekerk, D., van Heerden, C., Davel, M., Kleynhans, N., Kjartansson, O., Jansche, M., and Ha, L. (2017). Rapid development of TTS corpora for four South African languages. In *Proc. Interspeech 2017*, pages 2178–2182, Stockholm, Sweden.

Wibawa, J. A. E., Sarin, S., Li, C. F., Pipatsrisawat, K., Sodimana, K., Kjartansson, O., Gutkin, A., Jansche, M., and Ha, L. (2018). Building Open Javanese and Sundanese Corpora for Multilingual Text-to-Speech. In *Proc. LREC*, pages 1610–1614, 7-12 May 2018, Miyazaki, Japan.

Zhang, H., Sproat, R., Ng, A. H., Stahlberg, F., Peng, X., Gorman, K., and Roark, B. (2019). Neural models of text normalization for speech applications. *Computational Linguistics*, 45(2):293–337.

Proceedings, pages 95–96
Paris, UNESCO Headquarters, 5-6 December, 2019.

Red T Translator/Interpreter Incident Database

Maya Hess, Naomi Robbins
Red T
477 West 22 Street, New York, NY 10011
mhess@red-t.org, naomirobbins99@gmail.com

Abstract

Translators and interpreters (T/Is) in conflict zones and other high-risk settings are targeted by state and non-state actors alike. To mitigate this critical state of affairs, relevant policies must be implemented or improved. However, to drive meaningful policy change for the more vulnerable members of this largely invisible profession, accurate figures are needed. That is why Red T, a nonprofit organization advocating for the protection of T/Is at risk, is calling on governmental, intergovernmental and academic bodies to contribute data to the first database cataloguing incidents of T/I persecution, prosecution, imprisonment, abduction, torture and assassination.

Keywords: Translator, Interpreter, Database, Conflict Zone, Terrorism

Résumé

Les traducteurs et interprètes se trouvant dans des zones de conflit et autres environnements à hauts risques deviennent des cibles pour les acteurs étatiques et non étatiques. Pour atténuer cet état des choses critique, des politiques pertinentes doivent être mises en œuvre ou améliorées. Cependant, pour appuyer des changements constructifs en matière de politiques pour les membres plus vulnérables de ces professions largement invisibles, des chiffres précis sont nécessaires. C'est pour cette raison que Red T, une organisation à but non lucratif défendant la protection des traducteurs et interprètes à risque, invite les organes gouvernementaux, intergouvernementaux et académiques à contribuer leurs données à la première base de données cataloguant les incidents de persécution, poursuites judiciaires, emprisonnement, enlèvement, torture et assassinat à l'encontre de traducteurs et interprètes.

Mots-clés : traducteur, interprète, zone de conflit, terrorisme

1. Introduction

At a time when turmoil, warfare and mass migration afflict many parts of the world, a growing body of academic literature has established both the importance and the vulnerability of translators and interpreters (T/Is) in conflict and post-conflict situations. In a humanitarian response to the plight of these linguists, the nonprofit Red T was founded in 2010 with the mission of protecting T/Is in high-risk settings. These settings include war zones, detention centers, sites of political unrest, and terrorism-related venues, as well as countries in which translators of books, news items and other textual material deemed controversial are persecuted. Red T's stated vision is a world in which T/Is can work free from fear of persecution, prosecution, imprisonment, abduction, torture and assassination. To accomplish this, the nonprofit engages in a variety of policy-focused and educational initiatives designed to raise awareness across the world among governments, intergovernmental organizations and the public at large.[1] While many of the advocacy efforts have borne fruit, they were at times hampered by a lack of data. Seeking to address this dearth of facts and figures, Red T started gathering information on T/I incidents for the first database on this topic.

2. Challenges in Data Gathering

The challenges in gathering open-source data on a subgroup of a profession that, by default, is largely invisible are manifold. Prime among them is the nature of the settings and incidents, which for obvious reasons are not conducive to transparency. These include war zones in which civilian T/Is are targeted by insurgents due to their collaboration with foreign militaries, foreign correspondents and other foreign entities; the terrorism arena; nation states with restricted freedom of expression that unjustly arrest and prosecute T/Is; and instances where literary translators are the victims of aggravated assault or homicide. Another difficulty arises from casualty statistics released by governments that subsume the T/I category under the catch-all heading of locally employed civilians, as well as governments and private defense contractors that are tightlipped when approached for T/I figures. Additionally, names often pose a problem, whether because of misspellings, inconsistencies in transliteration, naming conventions, name variants, or the popularity of certain names–all of which create the likelihood of confusion–not to mention the failure to name T/Is at all.[2]

[1] Please see https://www.red-t.org

[2] Frequently, T/Is are not considered sufficiently important to warrant mention; in other instances, their anonymity is preserved to protect their identity.

3. Red T Pedagogical Module

3.1 Collaboration With Columbia University

The Institute for Comparative Literature and Society at Columbia University in New York City conducts the Global Language Justice Initiative (GLJ),[3] a seminar sponsored by the Andrew T. Mellon Foundation. As its name implies, the GLJ explores issues arising from the interrelationship of language and justice. Within the framework of this initiative, GLJ launched a strategic partnership with Red T. One of the results was a week-long pedagogical module to be incorporated into the syllabus of a related graduate-level course. The module was drafted by Alexandra Méndez, a GLJ Fellow, with the assistance of Red T. It introduces students to the Red T cause and different types of T/I rights violations. The students' assignment is to contribute to the Red T database by: digitally scouring multiple media outlets of a selected country, region or conflict zone; tracking one or two incidents of T/I persecution, prosecution, imprisonment, abduction, torture or assassination; and providing a visual or narrative analysis of the identified content. As stated in the syllabus, the module's purpose is "for students to learn about linguists' rights, data gathering and analysis in the digital sphere, as well as contribute actively to a project of advocacy for the protection of translators and interpreters in crisis settings" (Méndez, 2019). While the specific learning objectives vary with a given course's academic focus, students uniformly gain valuable insights into the humanitarian arena.

The pilot run of the module took place during the 2019 spring semester in the context of a graduate course entitled "Global Language Justice in the Digital Sphere." Professor Isabelle Zaugg supervised the module, assigned various relevant readings and hosted Red T CEO Maya Hess, who gave a presentation on the topic that was followed by a Q&A session. Several weeks later, students handed in their contributions to the database.

This preliminary run yielded valuable data but also revealed that more structure was needed to guide students' research in an unfamiliar field. In response, Red T drafted a detailed template that will be deployed as an integral part of the module in a second run. With those results in hand and any necessary adjustments made, the module will be made available to universities worldwide, with an initial targeting of institutions with translation and interpretation programs.[4]

The above methodology is beneficial for several reasons: For one, it permits the capture of T/I incidents reported in non-English languages for a more globally inclusive database. Having such comprehensive data will allow for a deeper perspective on the issue, which, in turn, will inform Red T's policy efforts and media outreach. Second, T/I students will become sensitized to the fate of their colleagues at risk, while students from other fields of study will not only learn about an area of rising humanitarian concern but acquire an understanding of the critical importance of the T/I role in high-risk settings.

4. Conclusion

Red T believes that effective linguist-friendly policies and enhanced legal mechanisms must be undergirded by robust data. Thus, to supplement contributions from academia, Red T is urging governments to respond to our requests for casualty figures to the greatest extent possible.[5]

In the spirit of "Protect Translators | Protect Interpreters | Protect the World," we hope that our call for data will be answered by governments across the globe and facilitated by intergovernmental bodies such as UNESCO.

5. Bibliographical References

Méndez, A. (2019). *Pedagogical module: Database for the protection of translators and interpreters worldwide*. Unpublished.

[3] See https://languagejustice.wordpress.com

[4] Universities and other parties interested in participating in the Red T T/I Incident Database project should contact mhess@red-t.org

[5] In the case of combat linguists, this request also extends to defense contractors and insurers.

Proceedings, pages 97–100
Paris, UNESCO Headquarters, 5-6 December, 2019.

Preserving and Developing Indigenous Languages in the South African Context

Justus C Roux, Stellenbosch University, Stellenbosch, South Africa, jcr@sun.ac.za,
Sonja E Bosch University of South Africa, Pretoria, South Africa, boschse@unisa.ac.za

Abstract

This article focuses on the development of infrastructures in the South African context that are dedicated towards the preservation and development of the eleven official languages of the country. It indicates to what extent the aims of this UNESCO conference are already being met by local activities. Specific attention is also paid to the role of language in the digital age.

Keywords: Infrastructures, African languages, digital resources

Abstrak

Hierdie artikel fokus op die ontwikkeling van infrastrukture binne die Suid-Arikaanse konteks wat toegespits is op die ontwikkeling en preservering van die elf amptelike tale van die land. Dit verwys spesifiek na die doelstellings van hierdie kongres van UNESCO en toon aan hoedanig plaaslike aktiwiteite alreeds vordering gemaak het. Spesifieke aandag word ook gegee aan die rol van taal in die digitale era.

1. Introduction

Given the main aims and objectives[1] of this conference, i.e.

(i) to identify recommendations on how to harness technology for the preservation, support and promotion of languages including lesser-used and Indigenous languages,
(ii) to provide access to information and knowledge to all language users and facilitate their inclusion and participation in building sustainable knowledge societies,
(iii) to promote the human rights and fundamental freedoms of all language users to access and create information and knowledge in language they best understand

this paper focuses on a relatively recent development in the South African context that implicitly addresses all of the above mentioned aims and objectives.

2. Background

South Africa has a population of 58,78 million[2] and eleven official languages that are used to a more or lesser extent across various regions of the country. These languages comprise four Nguni languages (isiZulu, isiXhosa, Siswati and isiNdebele), three Sotho languages (Setswana, Sesotho sa Leboa and Sesotho), Tshivenda, Xitsonga, Afrikaans and English. These African languages as well as Afrikaans are considered to be resource scarce languages (Moors et al., 2018) in a domain of digital communication.

At the turn of the century several academics lobbied national government to take the lead in setting up an infrastructure that could support the development of all official languages to keep these languages abreast of developments in the ICT and Human Language Technology (HLT) domains. This eventually led to the establishment of a Ministerial Advisory Panel in 2002 tasked to develop a strategic plan for the implementation of HLT in South Africa. In 2006 the South African Department of Arts and Culture established an HLT Unit responsible for driving the new HLT strategy which supported the research community with funds for collecting reusable resources and for developing appropriate NLP applications. One of the main recommendations of the Advisory Panel related to setting up a central repository to ensure the availability of reusable digital language resources for all official languages of the country. This recommendation eventually became a reality with the establishment of a Resource Management Agency (RMA) in 2012, amidst a growing number of research and development projects in HLT conducted by academics, see Roux and Ndinga-Koumba-Binza (2019). Through commissioned and own projects, the RMA rendered impressive results in the acquisition, enhancement and distribution of (South African) language resources and software tools. Within the first two years it had 258 registered users of which 157 were from South Africa, seven from five African countries, and 94 from 16 other countries worldwide with a total resource download at that point in time of 1 141. These resources and tools found their way into various research and development projects worldwide.

The RMA was eventually merged with the new South African Centre for Digital Language Resources (SADiLaR)[3] in 2017. The establishment of SADiLaR may

[1] https://en.iyil2019.org/events/lt4all-international-conference-on-language-technologies-for-all-enabling-linguistic-diversity-and-multilingualism-worldwide/ (accessed 13.11.2019)

[2] http://www.statssa.gov.za/ mid-year estimate 2019 (accessed 13.11.2019)

[3] http://www.sadilar.org (accessed 13.11.2019)

be regarded as fruition of numerous activities of researchers over a period of at least 15 to 20 years, eventually providing a long-term infrastructure in support of research and development. SADiLaR is a national entity sponsored by the Department of Science and Innovation (DSI) of the South African government, where it is a member of a new South African Research Infrastructure Roadmap (SARIR). This research and development entity comprises a Hub (hosted by the North-West University in Potchefstroom) linked to different Nodes such as the Department of African Languages at the University of Pretoria, the Department of African Languages at the University of South Africa (UNISA), the Centre for TexT Technology (CTexT) at the North-West University, the Council for Scientific and Industrial Research (CSIR) in Pretoria, and a consortium of universities (ICELDA). SADiLaR furthermore links up to similar international centres: currently it is a CLARIN C-centre and intends to apply for CLARIN B-centre status given the growth in activities. Furthermore, CLARIN, ELRA and the Linguistic Data Consortium (LDC) are all represented on the Scientific Advisory Board of SADiLaR, rendering valuable advice to the new entity.

Although the initial focus of SADiLaR is on the official languages of South Africa, it aspires to be of support to all major languages spoken in the sub-Saharan region. Researchers from these areas are invited to contact SADiLaR in this regard.

The main point to be made in this section relates to the establishment of appropriate infrastructures and calls for engagement with cultural organisations and governmental structures emphasising the shared responsibility towards the languages spoken in the particular country. It is necessary to keep as many indigenous languages alive in an ever-growing technological era, and to enable all speakers of indigenous languages, be they fully literate, semi-literate or even illiterate, to use their own languages for communicating both with other humans and with machines.

SADiLaR currently runs two programmes:

(i) A **digitisation programme**, which entails the systematic creation of relevant digital text, speech and multi-modal resources related to all official languages of South Africa. The development of appropriate natural language processing software tools for research and development purposes are included as part of the digitisation programme.

(ii) A **Digital Humanities programme**, which facilitates the building of research capacity by promoting and supporting the use of digital data and innovative methodological approaches within the Humanities and Social Sciences

The next three sections below will describe some of the activities of SADiLaR implicitly supporting the aims of this LT4All conference sponsored by UNESCO.

3. Harnessing technology

One of the prerequisites of 'harnessing' specific technologies relates to the availability of contents. In attempting to implement machine translation (MT) technologies it follows that ample digital text data of at least two languages should be available, firstly to train the system and then to run the translation engine, converting language A to language B. Similarly, the application of automatic speech recognition (ASR) technologies implies availability of different kinds of digitised speech depending on whose speech needs to be recognised. One of SADiLaR's aims as an enabling entity therefore is to accrue representative sets of digital text and speech corpora of the official languages of the country. Representative refers to language data acquired from as many different sources as possible. In the case of text this will imply newspaper items, literary works from different genres, instruction manuals, advertisements, formal government documents etc. In the case of speech, the data collected could be male or female speech, speech of young and old speakers, or speech regardless of age and gender, or the environment where speech is uttered, on a street corner with background noise, or in a motor vehicle listening to a GPS navigator.

The resource index[4] of SADiLaR already boasts a wide arrange of annotated text and speech resources as well as software tools which are freely available to researchers world-wide. Given these resources, and those still to be developed in effect function as living archives representing language types at different points in history and hence preserving these languages. At the same time the node at UNISA is highly involved in developing African Wordnets (AWN), an example of a multilingual knowledge resource covering all nine official African languages of South Africa. The open source nature of many wordnets (as is the case of the AWN as well) results in applications in different fields of research, making it "an ideal tool for disambiguation of meaning, semantic tagging and information retrieval" according to Morato et al. (2004:270). Concerning language learning applications, the use of wordnets to automatically generate vocabulary tests for second language acquisition has been reported on by Susanti et al. (2015). The usefulness of the current isiZulu Wordnet in a language learning application is also investigated by Bosch and Griesel (2018) who demonstrate how the unique sense identification features of wordnets can be incorporated into a language learning system thereby improving user interaction.

One of the nodes of SADiLaR, the Centre for Text Technology (CTexT) has made significant strides in developing machine translation technology for interactive translation between various official languages of South Africa. Software developed for translators includes the following tools[5] :

(i) "Integrated Translation Environment (ITE): The ITE is computer software that assists translators with their translation process, making use of translation

[4] https://repo.sadilar.org/handle/20.500.12185/1 (accessed 14.11.2019)

[5] http://humanities.nwu.ac.za/ctext/autshumato (accessed 20.11.2019)

memories and glossaries. It can be used to translate between any two natural language pairs in any of the South African languages."

(ii) "Translation Memory and Glossary Integration System (TMG): The TMG is a crowd-sourced platform from which translation resources can be provided as well as obtained. The sharing of translation resources between various translation units and freelance translators can ensure improved consistency and increase productivity throughout translation projects. These in turn can provide more access to information for everyone in their native languages. Users can rate and comment on resources, in order to give others an indication of the quality of a specific resource."

(iii) "Autshumato Terminology Management System (TMS): TMS helps with the development of terminology databases which contain terminology from different languages."

(iv) "Other resources that were developed as part of the Autshumato project include alignment software, a PDF extractor and a text anonymiser to safeguard privacy when using Autshumato ITE."

These cursory examples demonstrate how different technologies have been harnessed and applied to local languages by SADiLaR and two of its Nodes.

4. Access to information and knowledge

In its aim to provide access to information and knowledge, SADiLaR is currently involved in co-operation of the Dutch Language Union developing online Language Portal for two African languages, Setswana and isiXhosa. This two-year pilot project links up to similar portals developed for Dutch, Frisian and Afrikaans.[6] These two portals provide (in certain instances) new information through:

(i) a Grammar portal, comprising sub-domains such as Phonetics/Phonology, Morphology and Syntax,
(ii) a Dictionary portal,
(iii) a Corpus portal and
(iv) an Advice portal and forum.

This Advice portal and forum provides an infrastructure for "… language learners, language practitioners of various educational levels access to experts in the language to answer questions related to meaning, terminology and standardized spelling and grammar."[7] The web obviously provides multimodal content to be displayed, such as texts, speech and video which all contribute to new learning experiences.

5. Create information and knowledge

The nature of research in the Humanities and Social Sciences (HSS) has globally undergone a major paradigm shift over the last two decades due to:

(i) advances in the domain of Information Communication Technologies (ICT), and
(ii) increased access to digital resources.

This has given rise to an ever-expanding interdisciplinary domain of research and development, referred to as *Digital Humanities* (DH).

Given the development of large multilingual corpora, SADiLaR operates in close co-operation with the Digital Humanities Association of Southern Africa (DHASA).[8] These corpora can be searched by means of a wide range of applicable software, and as has been proven elsewhere, this more than often provides specific new knowledge. The strategic importance of SADiLaR became even more relevant as new methodological approaches toward research and development in the domain of Humanities and Social Sciences are posing new challenges to researchers. It was therefore strategically important to assist researchers not only in getting access to large corpora of authentic digital data and applicable software tools, but also to acquire skills related to the use of such data in order to render high quality research outputs nationally and internationally. This was furthermore a deliberate attempt to incubate the field of Digital Humanities in the South African context with benefits to society, academia, industry and government.

The close co-operation between SADiLaR and DHASA has led to a wide range of training programs such as for intance, three GROBID Dictionary Workshops[9] of SADiLaR across the country during 2018 related to "a machine learning infrastructure for creating structured lexicographic data from digitised dictionaries."

Given the growth of DH, SADiLaR has also become the academic home of the first professor in Digital Humanities in South Africa in 2019.

6. Conclusion

One of the first projects of SADiLaR was conducted by its node at the Council for Scientific and Industrial Research (CSIR) focusing on a national audit of HLT activities in South Africa (Moors et al., 2018:558). This audit was a follow up of a previous audit in 2009. One of the main findings was the following:

"Based on the comparisons between datasets and calculating the increase in resources, we were able to determine that there is an increase in resource availability for most South African languages. However, languages such as Xitsonga, Tshivenda, Sesotho, siSwati and isiNdebele still remain under-resourced. We were further able to deduce that more text than speech resources are currently available in South Africa. In addition to the comparison between resource types, we also determined the maturity and accessibility of the resources in all official languages in South Africa."

It appears that activities related to the preservation, and the development of indigenous languages of South Africa have

[6] Cf. www.taalportaal.org (accessed 14.07.2019)

[7] Quoted from original Project proposal: *Toward language portals for South African languages*.

[8] www.digitalhumanities.org.za (accessed 17.11.2019)

[9] http://digitalhumanities.org.za/index.php/dhasa-news (accessed 14.11.2019)

shown remarkable progress over the last decade. Implicitly supporting this view, Kaschula (2019:619) is of the opinion that "Technology is used here to capture, archive, and disseminate literary work in African languages (…). It is envisaged and suggested in this chapter that in future HLT will become the cornerstone of intellectualization of African languages.'

7. Bibliographical References

Abdullah, N. & Ibrahim, R., 2015, 'Managing information by utilizing WordNet as the database for semantic search engine', International Journal of Software Engineering and Its Applications 9(5), 193–204. https://doi.org/10.14257/ ijseia.2015.9.5.19

Bosch, Sonja and Griesel, Marissa. African Wordnet: facilitating language learning in African Languages. In: Francis Bond, Takayuki Kuribayashi, Christiane Fellbaum, Piek Vossen (eds) Proceedings of the 9th Global WordNet Conference (GWC 2018), 8-12 January, 2018, Nanyang Technological University (NTU), Singapore. Pp. 309-316.

Kaschula, R.H. and Nkomo, D. (2019). Intellectualization of African languages: Past, Present and Future, pages 601-622. The Cambridge Handbook of African Linguistics. H.E. Wolff (ed.) Cambridge. United Kingdom.

Moors, C., Calteaux K., Wilken, I and Gumede, T. 2018. Human language technology audit 2018: Analysing the development trends in resource availability in all South African languages. ACM International Conference Proceeding Series, pp. 296–304. Available from https://doi.org/10.1145/3278681.3278716 (Accessed on 2019-03-14)

Morato, J., Marzal, M.A., Lloréns, J. and Moreiro, J. 2004. WordNet Applications. In: Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): Global Wordnet Conference Proceedings 2004, pp. 270–278. Masaryk University, Brno.

Roux, J.C. and Ndinga-Koumba-Binza, H.S. (2019). African Languages and Human Language Technologies, pages 623-644. *The Cambridge Handbook of African Linguistics*. H.E. Wolff (ed.) Cambridge. United Kingdom.

Susanti, Y., Iida, R. & Tokunaga, T., 2015, 'Automatic generation of English vocabulary tests', in M. Helfert (ed.), Proceedings of the 7th International Conference on Computer Supported Education (CSEDU 2015), Lisbon, Portugal, 23–25 May 2015, pp. 77–78.

Proceedings, pages 101–103
Paris, UNESCO Headquarters, 5-6 December, 2019.

Be Not Like the Wind: Access to Language and Music Records, Next Steps

Nick Thieberger, Amanda Harris
University of Melbourne, University of Sydney
School of Languages and Linguistics, The University of Melbourne, Parkville, VIC 3010, Australia
PARADISEC, Sydney Conservatorium of Music, C41, University of Sydney, NSW 2006, Australia
thien@unimelb.org.au, amanda.harris@sydney.edu.au

Abstract

Language archives play an important role in keeping records of the world's languages safe. Accessible audio recordings held in archives can be used by speakers of small and endangered languages, and their communities, and provide a base for further research and documentation. There is an urgent need for historical analog tape recordings to be located and digitised, as they will soon be unplayable. PARADISEC holds records in 1228 languages. We run training for language documentation and are developing technologies to localise access to language records. A concerted effort is needed to support language archives and sustain language diversity.

Keywords: archives, language diversity, PARADISEC

Em i no olsem win – Painim tok peles na musik rekod
Wanpela kain ples olsem akaiv i save lukautim ol rekod or pepa bilong ol kain kain tok ples bai stap gut long bihain taim. Planti ol liklik tok ples ol klostu dai nau. Tok ples bilong ol manmeri na komuniti mas usim akaiv long helpim wok bilong painim aut moa na raitim ol pepa bilong ol tok ples. PARADISEC i gat 1228 tok ples. Mipela painim ol rikoding long taim tumbuna we ol tok ples i stap long keset tep bilong dijitaisim nau o sapos nogat bai ol bagarap. Taim nau long sapotim ol tok ples akaiv long lukautim planti kain kain tok ples.

1. Introduction

Me, selwan ag kupi em̃ae, tiawi itraus traus traus traus, natrauswen ga itaos nlag. Itrausi pan kaipa. Me komam uta laap kin uto mau, a? Malen umat, inom.
But when you are far away [and can't record him] the old man can talk and talk and talk, his story is like the wind. He tells it and it is gone. But there aren't many of us left. When we die, it will be finished.
†Kalfap̃un Mailei, 1998, Erakor Village, Efate, Vanuatu

In Melbourne recently a speaker of a language from Papua New Guinea looked through PARADISEC's webpage and we searched for the name of his language there. He was amazed to find recordings of his grandfather, never having expected to find anything in his language at all. For most of the world's small languages there is little or nothing available on the web, with most records, if they exist, still in analog form. In the passage quoted above, Kalfap̃un Mailei, an elderly man in 1998, was conscious of the need to record oral tradition so that it is available for others to hear in the future, and not, like the wind, here now and then gone. Language archives give us a glimpse of the richness of oral tradition, while they can never be a complete view of a language, these records of performances nevertheless provide both a cultural treasure for the speakers and their communities, and a research base for study of the world's languages.

Archives typically hold outputs of fieldwork, and so can have many hours of recordings for a language, which are often the only known recordings for that language. Of the 7,000 languages spoken in the world today, there are records of only a small proportion (Thieberger, 2016). Records of endangered languages that are unlikely to be spoken by a next generation of speakers, or have ceased to be spoken at all, are particularly valuable as they may be the only recorded source of information about that language. And getting good records back to this community can also help to strengthen the language, assisting in relearning older styles or performances. It also allows current speakers to enrich archive catalogs with their memory of what is recorded and its place in their society. There is a need for a concerted effort to index what is known for each language (see Thieberger 2106), as will be discussed below. An important first step is to locate and digitise all analog recordings, as we know that analog tape will not survive for much longer.

2. Language and music archives

The Open Languages Archive Community (OLAC) lists 60 language and music archives (http://www.language-archives.org/archives.php). 15 of these are represented by the Digital Endangered Languages and Musics Archives Network (DELAMAN). Each archive typically represents a particular geographical region. Community-agreed metadata standards enable harvesting by services like OLAC, which then creates a single page for each language that has a language code (ISO-639-3) thereby increasing the discoverability of language and music collections. Ideally, more fine-grained codes would be incorporated into such searches, for example, glottolog,[1] and more detailed regional codes, where they exist (like Austlang[2] in Australia).

Language archives provide this information as well as: licensing content; making content available; digitising analog materials; quality assurance; promotion of collections to source communities; conversion to archival

[1] https://glottolog.org

[2] https://collection.aiatsis.gov.au/austlang

and delivery formats; enforcing minimal descriptions of the items; providing citable forms of primary data.

The Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) was established in 2003 as a collaboration between linguists and musicologists at the University of Sydney, the University of Melbourne, and the Australian National University (Thieberger and Barwick, 2012). Critically, it was built by researchers who saw a need for a discipline-specific archive to deal with an inherited backlog of audio recordings that had no other prospect of being digitised.

3. What PARADISEC has done

The collection currently holds:
1228 languages
529 collections
26,395 items
287,511 files
7,397 users
11,500 hours
(At 8 November 2019)

Our catalog provides feeds that are harvested by external services, like the Open Language Archives Community, which increases the findability of an item in our collection. This means that even the most remote user who has internet access can find records.

We have received various awards and recognition for our work. In 2013, PARADISEC was added to the UNESCO Australian Memory of the World Register. In 2019 we received the World Data System[3] data seal, signifying we conform to all necessary standards.

Our initial motivation was preserving heritage recordings, and that continues to be an important part of our daily work. However, having built a relatively simple system for accessioning new items and collections, we are now also receiving numerous digital collections, some deposited in the course of fieldwork or soon after recordings were made. We are keen to help current fieldworkers to adopt methods that give them greater access to their own recordings, and, at the same time, make their collections 'archive-ready', reducing the amount of work required for their accession into our archive.

This leads to a focus on training in new methods so that the process of recording, transcription and annotation of transcripts all result in records that can be reused later on. We train academics and we train community members to do their own recording. We have encountered examples of recordings made on poor equipment, or where the microphone was too far from the speaker, so that little is audible. Low resolution recordings can be difficult to use for other purposes, like phonetic analysis or in creating teaching materials for the language. Transcripts made on paper can't be searched on a computer, and transcripts typed in a word-processor don't have timecodes that link back to the media. Current tools for transcription[4] insert timecodes for each chunk of the transcript and this means that this chunk can be played immediately and so allow you to cite down to the level of a word or sentence, strengthening research practice, and all the more so if the media is stored in a public online archive, like PARADISEC, and the media can be played directly from there. Articles referencing a story or a sentence can include a link for readers to follow to hear that item.

There is only a narrow window of time before analog tapes arising from historical field research become unplayable, both because they are on fragile media, and because of the increasing scarcity of playback machines. We are part of an international network of archives that is running an ongoing survey, called "Lost & Found" which asks for information about tape collections that need to be digitised. As a result of responses to this survey we have digitised fifteen collections of tapes. For example, we arranged for a collection of six hundred tapes from Madang in Papua New Guinea held at the Basel Museum (Switzerland) to be sent to our colleagues in the Netherlands for digitisation. A small collection of eight tapes in Yonggom (Papua New Guinea) were sent to our sister archive in the USA who digitised the tapes and sent the files to us to accession. Similarly, we arranged for a collection of 44 tapes in the Wampar (Papua New Guinea) language, recorded in the years between 1958 and 1972 and held on cassettes in Switzerland, to be digitised by another archive in London who then sent us the files.

While we work within the academy, we recognise that many of our colleagues do not take seriously the need to create lasting records in the course of their fieldwork, evidenced by the number of academic works published about languages over the past 30 years and the lack of archived records for those languages.[5] Accordingly, we make an effort to run regular training workshops and to advocate for the adoption of new methods that will increase the archivability of primary research data.

Further, there is an increasing amount of documentation being produced by speakers, some intentional, and some incidental to using social media. Both kinds of recordings risk being lost if there is nowhere for them to be housed, but social media is especially difficult to capture without a concerted effort. It is beyond our current ability to capture this, but it would be useful to have an automated service to recognise non-mainstream languages in social media, and then harvest that material into an archive.

4. Return of archival files

We have built a catalog that makes it relatively easy for material to be found in PARADISEC, assuming an internet connection and literacy in English. The normal kinds of search terms are provided: language, country, person, role, data, geography. To get the files to people with little or no internet access we have explored ways of sending copies of archival collections to source communities or nearby regional centres. An obvious way of doing this is to send all items for a given language or place on a hard disk to the local cultural centre or museum. This can work well, but also requires a catalog of those files to be created so that the contextual information in the catalog can be seen

[3] https://www.icsu-wds.org

[4] http://www.dynamicsoflanguage.edu.au/research/resources-for-linguistic-tools/

[5] see the analysis here: http://www.paradisec.org.au/blog/2016/07/finding-what-is-not-there/

together with the files. PARADISEC has a system that writes a text file (in XML) to the items in the collection each time the catalog entry is saved In this way, each item or set of files is self-describing so we can aggregate all of these text files for any given set of items and create a simple (html) catalog of just that collection.

But what about those places that don't have computers and so can't use a hard disk? We have built local wifi transmitters with hard disks that can be used in this situation[6]. The wifi transmitter is called a Raspberry Pi and costs less than AUD$100. It can be set up to transmit within a small area and so make this material accessible on mobile phones or tablets.

We are currently developing an OCFL[7] based version of our collection and that uses a similar principle of including the metadata with the object exported as RO-Crates[8]. The combination of RO-Crate described items stored in a standardised format could be a means to stepping into a modern, scaleable catalog application able to support many communities and many terabytes of data.

5. Collaboration with regional agencies to digitise their tapes

PARADISEC has established working relationships with agencies in our region like the Vanuatu Cultural Centre, Institute of Papua New Guinea Studies, University of French Polynesia, and the University of New Caledonia, among others. In 2014 we received funding to digitise hundreds of tapes held by the Solomon Islands Museum in Honiara. As we continue to run workshops in the Pacific region on issues around language recording methods, transcription, and data management, we continue to be offered tapes to digitise from local language authorities who no longer have the means to play tapes they created in the past. In November 2019 we collaborated in a seminar at the University of French Polynesia, the third we have run over the past few years. Present at the seminar were representatives of the various local language academies: Tahitian, Pa'umotu, and Marquesan. Earlier, we had worked to digitise 50 cassettes with some of these agencies, and at this event we received 19 cassettes, 6 minidisks, and 10 compact disks, entrusted to us to digitise and return. There are many such collections that we are yet to find, but each requires a relationship of trust with the owners, to let us take such important material to our offices where it can be properly assessed, cleaned, and digitised. In part, our longevity provides the status that allows us to build deeper relationships with these agencies and with the owners of other collections.

6. Conclusion

Without our work at PARADISEC over 11,500 hours of audio records would not have been preserved. However, we know there are many more records that have not yet been located by collecting agencies, and that there are not enough language and music archives in the world to deal with the quantity of material that has yet to be discovered. There needs to be a concerted effort to support new archives so that, in the words of Kalfaƀun Mailei, these words will not be like the wind, but will continue to enrich cultural practices, and help to show the value of the world's many small languages.

7. Acknowledgements

Thanks to Steven Gagau for translating the abstract for this paper into Tok Pisin. PARADISEC acknowledges funding support from the Endangered Archives Programme grant 693 (Preservation of Solomon Islands analogue recordings), Australian Research Council (ARC) LIEF program (2003, 2004, 2006, 2011), ARC Centre of Excellence for the Dynamics of Language, ARC Future Fellowship FT140100214.

8. Bibliographical References

Nick Thieberger. 2016. What remains to be done – Exposing invisible collections in the other 7000 languages and why it is a DH enterprise. In *Digital Scholarship in the Humanities*. 32(2), 1:423–434 . http://dx.doi.org/10.1093/llc/fqw006

Nick Thieberger and Linda Barwick. 2012. Keeping records of language diversity in Melanesia, the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC), pp. 239-253 in Nicholas Evans and Marian Klamer (Eds). *Melanesian languages on the edge of Asia: Challenges for the 21st Century*. [LD&C Special Publication No. 5]. Honolulu: University of Hawai'i Press. http://scholarspace.manoa.hawaii.edu/handle/10125/4567

[6] Discussed further here: http://www.paradisec.org.au/blog/2018/07/local-wifi-versions-of-paradisec/

[7] https://ocfl.io/)

[8] https://researchobject.github.io/ro-crate/

Proceedings, pages 104–107
Paris, UNESCO Headquarters, 5-6 December, 2019.

Machine Translation Literacy as a Social Responsibility

Lynne Bowker
School of Translation and Interpretation, University of Ottawa
70 Laurier Ave East, Hamelin Building (401), Ottawa, ON, K1N 6N5, Canada
lbowker@uottawa.ca

Abstract

Machine translation is easily accessible and easy to use, but this doesn't mean that everyone uses it in an informed way. We suggest that translators have a social responsibility for helping people outside the language professions to become informed users of machine translation, and that partnering with libraries could provide a means of reaching and educating a broad cross-section of citizens. We briefly summarize key elements of a machine translation literacy workshop that we piloted with two academic libraries, and we outline our plans for the next phase of the project with a public library.

Keywords: machine translation, machine translation literacy, social responsibility, ethic of care, libraries

Résumé

La traduction automatique est facilement accessible et facile à utiliser, mais cela ne signifie pas que tout le monde adopte une approche raisonnée. Nous suggérons que les traducteurs ont une responsabilité sociale de fournir une aide à ceux qui ne font pas partie des professions langagières pour qu'ils puissent devenir des utilisateurs critiques de la traduction automatique. De plus, nous suggérons qu'un partenariat avec les bibliothèques pourrait fournir un moyen d'éduquer un large éventail de citoyens. Nous résumons brièvement les éléments clés d'un atelier qui présente une approche raisonnée de la traduction automatique que nous avons mise à l'essai avec deux bibliothèques universitaires, et nous décrivons nos plans pour la prochaine étape du projet avec une bibliothèque publique.

1. Introduction

Although machine translation is nearly ubiquitous, not everyone is a critical user. We advocate for an ethic of care where translators can demonstrate social responsibility by helping those outside the language professions to become informed users of machine translation. We explain how partnering with libraries could allow translators to educate a broad cross-section of society, and we outline key elements in a machine translation literacy workshop that we piloted with two academic libraries. Finally, we briefly note our plans for the next phase of the project with a public library.

2. Machine Translation: An Evolving Landscape

The context of machine translation use has evolved considerably since the first tools were developed in the late 1940s and early 1950s. This includes a change in the types of user, in how the tools work, and in the type of education needed to ensure critical use.

2.1 Machine Translation is "In the Wild"

Until relatively recently, machine translation tools could be found primarily in the hands of researchers or language professionals. Then in 2006, Google launched a free, online machine translation system called Google Translate. Other companies followed suit with their own browser-based machine translation tools, including Bing Microsoft Translator, DeepL Translator, Baidu Translator, and Yandex.Translate, to name a few. In addition, some of these translation engines provide built-in translation options in tools such as the Google Chrome browser, or for platforms such as Facebook or Twitter. We could say that machine translation is now "in the wild", meaning that these tools are no longer restricted to language professionals but are available to everyone with an internet connection. Non-language professionals are using machine translation to assist with tasks such as conducting genealogical research (Vestal, 2016) or searching for international patents (Nurminen, 2019), among other uses.

2.2 Machine Translation is Easy to Use

Some translation technologies, such as translation memory systems, are still used primarily by language professionals. These tools can be quite sophisticated and require specialized training to learn *how* to use them. In other words, they can be complicated to use from a technical standpoint – knowing which file to open, which option to select, which filter to apply, and so on. In contrast, browser-based machine translation or built-in machine translation tools are very simple to use. Sometimes it takes just one click! However, the effortlessness with which we can employ these tools means that it is very easy to use them in an unthinking or non-critical way, which could lead to problems.

2.3 Machine Translation is Undergoing a(nother) Paradigm Shift

Another change that has occurred is that the underlying approach to machine translation has changed. Machine translation research began just after the Second World War. For approximately 50 years, the main approach to machine translation was known as Rule-Based Machine Translation (RBMT) (Hutchins and Somers, 1992). With RBMT, developers approached machine translation in a way that was similar to how linguists study language – through grammar rules and bilingual lexicons. These systems had limited success, and a common problem was that of "translationese", where the translated text would be awkward or overly literal. Around the turn of the millennium, researchers began to adopt a corpus-based or data-driven approach to machine translation, where statistics rather than linguistics took centre stage (Koehn,

2010). Statistical machine translation (SMT) approaches allowed computers to do what they excel at: number crunching and pattern matching. With SMT, translation quality got noticeably better, and it was during this period that the previously mentioned free, online machine translation systems first began to appear.

In late 2016, the underlying approach to machine translation changed again. Still data-driven, today's state-of-the-art machine translation systems use artificial neural networks, coupled with a technique known as machine learning (Forcada, 2017; Way, 2019). Developers "train" neural machine translation (NMT) systems by feeding them enormous parallel corpora that contain millions of pages of previously translated texts. NMT systems use these examples to "learn" how to translate new texts. With this latest paradigm shift, the quality of machine translation output has further improved. If the texts produced by RBMT systems were often laughable, the output of NMT systems, though not perfect, may be quite usable for many purposes. However, users must show good judgement. For instance, Castilho et al. (2017) found that found that NMT systems often produce text that is more fluent and contains fewer telltale errors such as incorrect word order or other forms of "translationese". However, just because the NMT output reads well doesn't always mean that it's accurate or right for a user's needs.

3. Machine Translation Literacy

Just because machine translation is easily accessible, easy to use, and produces a quality of output that is reasonable for some purposes, this doesn't mean that we instinctively know how to optimize it or even to use it wisely in a given context. The need for a new type of digital literacy is emerging, which we refer to as machine translation literacy (Bowker & Buitrago Ciro, 2019).

Martin (2006) describes digital literacy as the awareness, attitude and ability of individuals to appropriately use digital tools to identify, access, manage, integrate, evaluate, analyze and synthesize digital resources, construct new knowledge, communicate with others, and to reflect upon this process. This definition emphasizes that critical thinking, rather than technical competence, is the core skill of digital literacy. Like digital literacy, machine translation literacy is primarily a cognitive issue, rather than a techno-procedural one. Using machine translation is easy; using it *critically* requires some thought. When faced with free, online machine translation, the important question is not *how to* but rather *whether*, *when*, and *why* to use this technology. With regard to *how*, we could more usefully frame this as 'how can users interact with this tool in order to improve the usefulness of its output?' By asking ourselves such questions, we can become informed and critical users of machine translation tools, rather than being people who simply copy, paste, or click without a second thought.

4. Machine Translation Literacy as a Social Responsibility

Translator education programs typically incorporate translation technology training into their curricula, and professional translators' associations also offer options for technology-related professional development. Therefore, we can be hopeful that language professionals are (becoming) machine translation literate, or at least have the means to do so. However, it is not clear how the many people outside the language professions can learn to become savvy machine translation users. We would like to suggest that translators have a social responsibility in this regard.

Drugan and Tipton (2017) recently observed that relatively little attention has been paid to the question of social responsibility in relation to translation, prompting them to propose a Special Issue of the journal *The Translator* on the topic of translation, ethics and social responsibility. In it, Drugan (2017: 128) notes "we understand social responsibility as individuals' responsibility to the wider society in which they live; that is, interpreters' and translators' responsibility to the broader social context beyond the immediate translated encounter".

In addition to this special issue journal, we can observe some other ways in which the language professions are beginning to engage with social responsibility. For instance, the translation profession is generally regarded as being a caring profession where volunteerism is widespread. As described by Federici and O'Brien (2019), translation can play a key role in reducing risk in crisis situations (e.g. following disasters such as cyclones or earthquakes, or during the spread of infectious diseases), and there are several organizations (e.g. Translators Without Borders, *Solidarités International*) that use volunteer translators to help address humanitarian needs in collaboration with various non-governmental organizations (NGOs).

Meanwhile, Cheung (2017) proposes that *plain language* can be used for social good and indicates that technical communicators who use plain language are exercising social responsibility. She argues that marginalized populations (i.e., people who are oppressed for any reason) have a lot of worries to occupy their minds. The greater stress and mental burden that marginalized populations experience can leave less working memory available for tasks such as reading and learning. Cheung (2017: 448) states "Using plain language to reduce cognitive load can be considered a political act that increases marginalized populations' opportunities to understand." She thus presents the use of plain language by technical communicators as an ethical imperative.

4.1 Towards an Ethic of Care

We suggest that an ethic of care, such as that put forward by Noddings (1984; 2002), presents a good framework for encouraging members of the language professions to promote machine translation literacy to those outside these professions. To date, if professional translators discuss machine translation with non-translators, it has tended to take the form of warning them off using this technology. For example, as outlined in Bowker (2019) the website of the Canadian Translators, Terminologists and Interpreters Council (CTTIC) contained a message actively dissuading people from using machine translation and warning of the dangers of relying on machine translated output. The message on the CTTIC site noted:

> As part of their mandate, CTTIC's member organizations have a duty to ensure the protection of the public. As such, CTTIC and its members urge users to exercise the highest degree of caution, and to call upon a certified translator for all their translation requirements.

This approach does not seem particularly helpful. It is not realistic to think that all people who seek translation services can afford to hire professional translators, nor is it likely that all translation jobs require a professional level of quality. If translators truly see themselves as "having a duty to ensure the protection of the public", perhaps they could carry out a greater service to society by helping potential users to become machine translation literate instead of trying to convince them to steer clear of machine translation altogether.

In her early work on the ethics of care, Noddings (1984) distinguishes between 'caring for' and 'caring about', but she initially brushes aside 'caring about', noting that it is too easy and involves a sort of benign neglect:

> I can 'care about' the starving children of Cambodia, send five dollars to hunger relief, and feel somewhat satisfied. I do not even know if my money went for food, or guns, or a new Cadillac for some politician. This is a poor second-cousin to caring. 'Caring about' always involves a certain benign neglect. One is attentive just so far. One assents with just so much enthusiasm. One acknowledges. One affirms. One contributes five dollars and goes on to other things. (Noddings, 1984: 112)

However, she later revisited this decision, noting that while the basic distinction between 'caring for' and 'caring about' remains important, the concept of 'caring about' actually does warrant more attention. Indeed, her later work, she puts forward the idea that 'caring about' provides a link between caring and justice:

> … we learn first what it means to be cared for. Then, gradually, we learn both to care for and, by extension, to care about others. This caring-about is almost certainly the foundation for our sense of justice. (Noddings, 2002: 22)

Noddings (2002) explains that 'caring about' moves us from the face-to-face world into the wider public world, where we are moved by compassion for others' suffering, we regret that they do not experience being cared for, and we are outraged when they are exploited. In cases where we cannot directly care for others, we express our care in other ways, such as by donating to charities, supporting certain social groups, or voting.

Noddings (2002) is careful to point out that 'caring about' presents some inherent flaws. For instance, at its worst, it can become self-righteous or politically correct, it can encourage dependence on abstractions, and it can elevate itself above 'caring for' others. Nonetheless, Noddings (2002) believes that 'caring about' (i.e., a sense of justice) is instrumental in in establishing the conditions under which 'caring for' can flourish. In other words, although the preferred form of caring is 'caring for', 'caring about' can help to establish, maintain and enhance it.

In this vein, if translators care about their fellow citizens, they could show this by using their expertise to help others become more informed about machine translation. In this way, people will be in a position to decide whether or not this technology meets their needs for a given task, and if so, how they can use it effectively in a critical way.

5. Resources and Infrastructure

Since translators or their associations may lack a suitable infrastructure and resources to deliver machine translation literacy instruction, an option may be to form partnerships with different types of libraries. One reason for partnering with libraries is that they are typically cross-cutting units that reach a wide cross-section of the populations that they serve. For instance, an academic library serves the entire range of disciplines covered by its host institution, and it offer services to students, staff and faculty alike. Meanwhile, a public library cuts across socio-economic classes, offering services to all members of the public. In addition, both academic and public librarians are already charged with delivering other types of literacy instruction, including information literacy, media literacy and digital literacy (e.g., Julien, 2005). This experience makes librarians well equipped to partner with language professionals in order to offer machine translation literacy training as part of their programming.

5.1 Machine Translation Literacy Instruction in Academic Libraries

In autumn 2019, we conducted a pilot project where we delivered machine translation literacy workshops to international university students, faculty, and staff in collaboration with two university libraries in Canada: Concordia University Library in Montreal and the University of Ottawa Library in Ottawa. We ran three workshops at Concordia and two in Ottawa with a combined total of over 100 participants.

Examples of the type of information that was shared with workshop participants include suggestions such as:

a) **Don't enter sensitive material into an online machine translation system.** Information that you type or paste into a free online machine translation system doesn't simply "disappear" when you close the window. Instead, the companies that own the machine translation system (e.g. Google, Microsoft) could keep the data and use it for other purposes.

b) **Be sure to cite and reference ideas, even if you translate the words.** Academic integrity must be respected even when using machine translation tools to translate ideas into another language.

c) **Try more than one machine translation system.** Today's state-of-the-art neural machine translation systems use large corpora of previously translated texts as examples to "learn" how to translate new texts. Keep in mind that each machine translation system is trained using different texts, so each system might "learn"

different things. If one system doesn't provide helpful information, then try another one. Also, remember that these machine translation systems are constantly learning. If a particular system doesn't meet your needs today, try it again next month and you could get different results.

d) **Consider the purpose of the translation.** Machine translation may be more useful or less useful for different types of tasks or texts. If you are using the translation simply to help you *understand* a text that has been written in another language, such as reading a research article as part of a literature survey for your thesis, then, a machine translation system can probably be quite useful for helping you to get the gist or the main message of that text. However, if you're planning to use machine translation to help you *write* a text (e.g., a term paper or an article for publication), then be aware that unedited machine translated text is *not* likely going to be of a high enough quality for such purposes. The machine translation output will need to be edited to improve the quality.

e) **Improve the output by changing the input.** You might have heard the expression "garbage in, garbage out"? Well, if you want to use a machine translation system to help you produce a good translation, the best thing that you can do is to write the input text in a clear and easy-to-read way. We call this "translation-friendly writing", and it includes things such as using short sentences, avoiding humour, idiomatic expressions or culture-bound references, and using full forms instead of abbreviated forms.

Participants were surveyed about the workshops. The vast majority of participants indicated that they had learned new things. Most said that they plan to integrate machine translation more regularly into their scholarly work, and that they now feel equipped to do so more effectively. More than half the respondents replied that they would be interested in taking a more advanced follow-up workshop.

5.2 Next Steps: Working with Public Libraries

Following on from the success of introducing machine translation literacy workshops in academic libraries, we now plan to expand this type of training to a broader public. To this end, we are currently working with the Ottawa Public Library to explore how machine translation literacy instruction can be usefully adapted for different types of public library patrons. In particular, members of the newcomer or immigrant community are interested in learning how to become more informed and critical users of machine translation. Other possible target audiences include teens and seniors. We aim to pilot workshops that have been adapted for these groups in 2020.

6. Conclusion

Machine translation is being increasingly used in our society, where it has the potential to help if used critically, but to harm if used carelessly. Translators bear some responsibility for helping those outside the language professions to become informed users of this technology. Partnering with libraries can provide a means of reaching and educating a wide range of machine translation users.

7. Acknowledgements

This project was first developed as part of the Researcher-in-Residence program at Concordia University Library.

8. Bibliographical References

Bowker, L. (2019). Fit-for-Purpose Translation. In M. O'Hagan, editor, *Routledge Handbook of Translation and Technology*, pages 453–468, London, Routledge.

Bowker, L. and Buitrago Ciro, J. (2019). *Machine Translation and Global Research: Towards Improved Machine Translation Literacy in the Scholarly Community*. Emerald Publishers, Bingley, UK.

Canadian Translators, Terminologists and Interpreters Council (CTTIC). Homepage. Available online: www.cttic.org/ [last accessed 3 October 2018]

Castilho, S. et al. (2017). A Comparative Quality Evaluation of PBSMT and NMT Using Professional Translators. In *Proceedings of Machine Translation Summit XVI*, vol. 1, pages 116–131, Nagoya, Sept.

Cheung, I. (2017). Plain Language to Minimize Cognitive Load: A Social Justice Perspective. *IEEE Transactions on Professional Communication* 60(4): 448–457.

Drugan, J. (2017). Ethics and social responsibility in practice: interpreters and translators engaging with an beyond the professions. *The Translator* 23(2): 126–142.

Drugan, J. and Tipton, R. (2017). Translation, ethics and social responsibility. *The Translator* 23(2): 119–125.

Federici, F. M. and O'Brien, S. (2019). Cascading Crises: Translation as Risk Reduction. In F. M. Federici and S. O'Brien, editors, *Translation in Cascading Crises*, pages 1–22, Routledge, London.

Forcada, M. L. (2017). Making Sense of Neural Machine Translation. *Translation Spaces* 6(2): 291–309.

Hutchins, W. J. and Somers, H.L. (1992). An Introduction to Machine Translation. Academic Press, London.

Julien, H. (2005). Education for Information Literacy Instruction: A Global Perspective. *Journal of Education for Library and Information Science* 46(3): 210–216.

Koehn, P. (2010). Statistical Machine Translation, Cambridge University Press, Cambridge.

Martin, A. (2006). Literacies for the digital age. In A. Martin and D. Madigan, editors, *Digital Literacies for Learning*, pages 3-25, Facet Publications, London.

Noddings, N. (1984). *Caring: A Feminine Approach to Ethics and Moral Education*. University of California Press, Berkeley.

Noddings, N. (2002). *Starting at Home: Caring and Social Policy*. University of California Press, Berkeley.

Nurminen, M. (2019). Decision-making, Risk and Gist Machine Translation in the Work of Patent Professionals. *Proceedings of the 8th Workshop on Patent and Scientific Literature Translation*, pages 32–42, Dublin, August.

Vestal, P. (2016) Why I love Google Translate. *Assoc. of Professional Genealogists Quarterly* 31: 84–86.

Way, A. (2019). Machine Translation: Where are we at today? In E. Angelone, M. Ehrensberger-Dow and G. Massey, editors, *The Bloomsbury Companion to Language Industry Studies*, pages 311–332, Bloomsbury, London.

Proceedings, pages 108–111
Paris, UNESCO Headquarters, 5-6 December, 2019.

The Endangered Languages Project (ELP): Collaborative Infrastructure and Knowledge-Sharing to Support Indigenous and Endangered Languages

Anna Belew
Endangered Languages Project
1890 East-West Rd., #569, Honolulu, HI 96822
anna@endangeredlanguages.com

Abstract

While global language loss is currently occurring at a rate higher than ever in known history, the prevalence of language documentation and revitalization work is also rising dramatically. However, many such initiatives are taking place in relative isolation, without clear pathways for collaboration and knowledge-sharing across nations and global regions. The Endangered Languages Project (ELP) addresses this need by providing infrastructure for the dissemination of knowledge about Indigenous and endangered languages, their vitality, and their revitalization efforts; sharing of digital resources related to Indigenous and endangered languages; and fostering international collaboration and networking between language workers and researchers.

Keywords: knowledge-sharing, infrastructure, collaboration

摘要

儘管現今全球的語言正在以史上最快的速度消失，語言保存和復振也在快速的進當中。然而，這些活動往往單獨的在進行，而且顯少有跨國或全球性的合作及知識途徑。瀕危語言計畫（Endangered Languages Project (ELP)）正是為了這個需求誕生，提供傳播原住民和瀕危語言知識、活力以及成果的設施，並與大眾分享原住民族和瀕危語言的數位資源，積極促進語言工作者和研究者之間的國際合作和人際網絡。

1. Global Language Endangerment and Revitalization

Of the roughly 7,000 languages spoken on Earth today (Eberhard, Simons and Fennig, 2019), nearly half—approximately 3,170 languages, or c. 45%—are currently at risk of falling silent (*Catalogue of Endangered Languages,* 2019). Another 183 languages, at least, are known to have fallen silent in the past half-century, and an additional 70 or more languages are currently being revived after having lost their last fluent speakers (ibid.). The majority of these are Indigenous and/or minoritized languages, and shift away from these languages often corresponds to other forms of socio-economic, political, and cultural marginalization and injustice.

While rates of global language loss are currently higher than at any point in known history, the incidence of language documentation and revitalization is also at an all-time high (Belew and Simpson, 2018). Indeed, Pérez Báez, Vogel, and Patolo (2019) report that, of the revitalization programs surveyed in their research, 65% were initiated after the year 2000, and 30% began in the current decade. At least 41 known revitalization programs have begun in the past ten years, and new initiatives are continually being launched (ibid.). Similarly, Thieberger (2017) reports that language documentation, if assessed in terms of grammars produced for small and endangered languages, has increased notably since the year 2000.

As programs to support linguistic diversity proliferate in all parts of the world, there is a need for digital infrastructure to support these initiatives with reliable data, technological tools, knowledge-sharing mechanisms, and avenues for collaboration. The following sections outline these needs, and how the Endangered Languages Project's digital services and resources can support those working to sustain Indigenous and endangered languages.

2. Needs for Infrastructure to Support Language Documentation and Revitalization

While the number of initiatives to document and revitalize the world's languages is growing, many language workers today are still carrying out their work in relative isolation, without a clear path for connecting with people and programs in other parts of the world. Language workers may be discouraged by the feeling carrying out language work alone, if there is limited support from populations and institutions in a given context. For example, one language activist from South Africa, in conversation with the author during a conference, described being demoralized by feeling that they were the only person engaged in this type of work; they had few connections with other language activists, and limited knowledge of initiatives in other parts of the world. However, after being invited by ELP to attend an Indigenous language revitalization conference to network, collaborate, and share knowledge with language workers from around the world, and maintaining online contact with their new colleagues after they returned to their home countries, the language activist reported that their work was reinvigorated by new motivation, ideas, and methods. Knowledge-sharing and networking with fellow language workers is often key to sustainable, effective language work.

Similarly, many researchers working with Indigenous and endangered languages lack a place to make their research more widely accessible, and to connect with

language stakeholders outside of academia. Moreover, reliable data on endangered languages may be difficult to come by for many researchers and language workers—it may be locked behind a paywall, buried in dense academic volumes, or written in a language not known to the reader—and without accurate information about a language and its current context, it is difficult to devise an effective strategy for documentation or revitalization. Similarly, policymakers, funding agencies, activists, and governments are in need of accurate data about each of the world's endangered languages, in order to make informed decisions about language policy, public outreach, and the allocation of limited resources (Hauk and Heaton 2018). As some prominent sources of language information are now moving to paid access models[1], and others are falling out of date after losing funding, there is an enormous need for online resources which provide free, up-to-date information about endangered and Indigenous languages.

In addition, helping stakeholders stay informed of global activity in the field of Indigenous and endangered languages is a significant need—as demonstrated by the spread of the highly successful language nest model from Aotearoa/New Zealand, to Hawai'i, to North America and Europe (Hinton 2018), the dissemination of effective ideas and best practices across the globe is crucial to ensuring that Indigenous and endangered languages can thrive in the coming decades.

Finally, as technology tools—such as smartphone keyboards, software localization, and language learning apps—become an increasingly central part of language maintenance and revitalization, there is a growing need for language workers to have access to best practice guides, networks of fellow practitioners, and reliable information to support the development of these technologies.

3. The Endangered Languages Project (ELP)

The Endangered Languages Project (ELP) is an online resource (accessible at www.endangeredlanguages.com) designed to promote and facilitate the documentation and revitalization of at-risk languages around the world. ELP aims to bring together language workers, researchers, and the public to share knowledge, best practices, ideas, and news about the world's Indigenous and endangered languages. The project serves both as a resource for information on the endangered languages of the world, and as an arena to collaborate with others working to document, revitalize, and promote Indigenous and endangered languages.

Launched in 2012, ELP was originally founded by four core partners: the First Peoples' Cultural Council (FPCC), the University of Hawai'i at Mānoa (UHM), The LINGUIST List at Eastern Michigan University, and Google.org. It is currently led by two founding partners (FPCC and UHM), and a Governance Council of individuals from nine organizations and universities around the world[2]. The website interface is currently available in seven languages (English, Spanish, French, German, Traditional Mandarin, Russian, and Brazilian Portuguese), and is currently in the process of being localized into additional languages, including Korean, Japanese, Italian, and Hindi. ELP plans to expand the site interface's availability in other languages, particularly Indigenous languages, as the project continues. The site currently has over 20,000 registered users, and receives an average of three million annual pageviews.

The following sections outline the services offered by ELP, and the ways in which ELP's infrastructure may be used and expanded to meet the needs of Indigenous and endangered languages online.

3.1 The *Catalogue of Endangered Languages* (ELCat)

The *Catalogue of Endangered Languages* (ELCat) is a database of information about the world's endangered languages. ELCat is a cost-free resource for language workers, researchers, educators, language communities, policymakers, funding agencies, and the public to learn about global language endangerment and revitalization.

ELCat provides information about the vitality and context of each of the world's endangered, dormant, and awakening[3] languages, which at the time of writing numbered 3,426. The ELCat database aggregates information from all available, reputable sources, including journals, books, field notes, reports from Indigenous language organizations, and direct communications from individuals working directly with a given language. To ensure the reliability of ELCat data, this information is vetted by the project's International Board of Directors, who are individuals with expertise in the languages of specific global regions; see (Campbell and Belew 2018) for more on ELCat's data collection practices. ELCat uses an original metric for assessing language vitality, the Language Endangerment Index (LEI), and provides visualizations of the vitality of all of the world's endangered languages in the form of a global map, as shown in Figure 1. For more on the LEI, see (Lee and Van Way 2016).

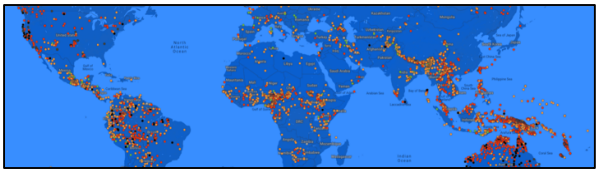


Figure 1: Partial view of ELP map of endangered languages, color-coded by vitality level.

[1] https://linguistlist.org/issues/30/30-4081.html

[2] http://endangeredlanguages.com/about/

[3] "Dormant" refers to languages which have lost their last fluent speaker in roughly the past 50 years, while the "Awakening" label is applied to languages which at some point had no fluent speakers, but are currently undergoing concerted revival efforts. See (Belew and Simpson 2018) for more on these languages.

3.2 Dissemination of Digital Multimedia Resources

The ELP website also offers a platform for users to disseminate digital multimedia resources—videos, audio files, images, documents, and web links—related to endangered and Indigenous languages. In addition, there is a discussion function for users to comment on, share information about, and make recommendations for resources similar to those included in the site. ELP allows all site users to upload multimedia resources, encouraging the free exchange of ideas and media on the topic of endangered languages. Resources uploaded to the ELP site are lightly moderated by the project's staff to ensure that inappropriate or irrelevant content is removed.

Users can browse multimedia resources by language—each language page features a "Resources" tab, displaying all relevant materials—as well as by category and keyword. Resource categories include topics such as language education, language and technology, and language advocacy and awareness; these categories aggregate topically similar materials from many different languages. For example, a user interested in language nests could browse all materials tagged "language nest" within the "language education" category, and access videos, documents, and news articles related to language nests worldwide. They can also view the profiles of other users who have submitted (or commented on) resources related to language nests, and connect with one another for collaboration and networking.

There are currently over 7,000 multimedia resources available on the ELP site, with new items being uploaded daily by users. Such a platform provides a useful avenue for resource discovery and dissemination, a key need in supporting the presence and use of endangered languages online.

3.3 Digital Training in Language Documentation

ELP seeks to address the need for free training and information related to language documentation, particularly among speakers of endangered and under-documented languages who do not have access to formal training programs, such as university courses or workshops, ELP has partnered with the Language Documentation Training Center (LDTC) at the University of Hawai‘i to offer digital training in language documentation. These eight-week webinars, held weekly via Facebook Live, provide training in basic facets of language documentation (audio/video recording, descriptive phonology, orthography development, etc.), as well as practice exercises and further reading. In addition to watching the livestreamed lessons, asking questions of the workshop leaders, and discussing the material with other learners, participants are also paired with "mentors" who have advanced skills in language documentation for more personalized guidance.

The first webinar, held in early 2018, drew 465 participants from 62 countries. Participants were particularly enthusiastic about the opportunity to network with language workers and researchers in other parts of the world; as discussed in §2, a sense of isolation in the work of language documentation, revitalization, and advocacy can be demoralizing, while collaboration and interaction can renew motivation and sustainability of documentation efforts. This type of interactive webinar is one method for both increasing access to language documentation training, and for facilitating networking and knowledge-sharing among speakers of Indigenous and endangered languages.

3.4 Revitalization Directory and Helpdesk

3.4.1 Revitalization Directory

ELP is currently developing a directory of revitalization initiatives around the world, based on the Global Survey of Revitalization Efforts developed by Pérez Báez, Vogel, and Patolo (2019). ELP's revitalization directory will build upon the results of Pérez Báez et al. and provide an expanded listing of language revitalization programs. Users will be able to submit information about revitalization programs they are involved with, such that the directory will constantly be expanded and updated.

The directory will allow ELP users to network with other revitalization workers who are working with similar methods, languages, or contexts. For example, a teacher at a Hawaiian-immersion high school could search for other secondary-education revitalization programs in order to find lesson planning ideas, or learn about successful revitalization methods for languages in similar sociopolitical contexts. In addition, the directory will highlight the growing prevalence and diversity of language revitalization work around the world; it is our hope that this will provide encouragement and a sense of global solidarity to language revitalization workers, and demonstrate the resilience of Indigenous and endangered languages in the face of pessimistic discourses of language loss. This directory will provide online infrastructure for knowledge-sharing and collaboration across global regions, while also serving as a reliable source of information on global trends in language revitalization, paralleling how ELCat provides data on global trends in language vitality (see Campbell and Okura 2018).

3.4.2 Revitalization Helpdesk

The ELP Revitalization Helpdesk will complement the directory of revitalization initiatives by providing case studies, best practices, and personalized human assistance for revitalization workers. It can be intimidating to attempt to begin a revitalization effort, particularly in the absence of connections with established programs or language workers. There are a myriad of questions to address—what methods exist, and what goals do they serve? What types of activities have proven effective in similar contexts? What kinds of resources are needed? etc.—and seeking answers in the academic literature can be daunting for laypeople.

The Revitalization Helpdesk will thus provide resources for strategic language revitalization planning, such as "roadmaps" and tools for assessment of a language's vitality and revitalization goals, drawing upon the extensive experience of ELP's founding partners at the First Peoples' Cultural Council. The Revitalization Helpdesk will also provide best-practice guides, profiles of

successful revitalization efforts, discussion areas to connect with other users engaged in language revitalization work, and a "helpdesk" which will connect language workers with volunteer experts in particular revitalization methods, language families, or geographic areas.

The directory and helpdesk are currently in development, and will launch in 2020.

4. Conclusion

The Endangered Languages Project provides infrastructure to support key aspects of language documentation and revitalization globally, including a free, reliable database of information about endangered languages and revitalization initiatives; online dissemination of multimedia resources; internet-based training in language documentation; and a "helpdesk" for language revitalization workers to find guidance, ideas, best practices, and collaboration with other revitalization initiatives around the world.

9. Acknowledgements

The *Catalogue of Endangered Languages* and Endangered Languages Project have received support from the U.S. National Science Foundation (grants BCS-1058096 and BCS-1057725), the Henry Luce Foundation, Google.org, the University of Hawai'i at Mānoa, and the First Peoples' Cultural Foundation. Translation of this paper's abstract into Mandarin was provided by Khái-īn Lîm.

5. Bibliographical References

Belew, A. and Simpson, S. (2018). Language extinction then and now. In Lyle Campbell and Anna Belew, editors, *Cataloguing the World's Endangered Languages*, pages 49–65. Routledge, New York, 1st edition.

Campbell, L. and Belew, A. (2018). Cataloguing the world's endangered languages. Routledge, New York, 1st edition.

Campbell, L. and Okura, E. (2018). New knowledge produced by the *Catalogue of Endangered Languages*. In Lyle Campbell and Anna Belew, editors, *Cataloguing the World's Endangered Languages*, pages 79–84. Routledge, New York, 1st edition.

Catalogue of Endangered Languages (ELCat). (2019). University of Hawai'i at Mānoa, available online: http://www.endangeredlanguages.com.

Eberhard, D.M., Simons, G., and Fennig, C. (2019). Ethnologue: Languages of the world. 22nd edition. SIL International, Dallas, available online: http://www.ethnologue.com.

Hauk, B. and Heaton, R. (2018). Triage: Setting priorities for endangered language research. In Lyle Campbell and Anna Belew, editors, *Cataloguing the World's Endangered Languages*, pages 259–304. Routledge, New York, 1st edition.

Hinton, L. (2018). Approaches to and strategies for language revitalization. In Kenneth L. Rehg and Lyle Campbell, editors, *The Oxford Handbook of Endangered Languages*, pages 443–465. Oxford University Press, Oxford, 1st edition.

Lee, N.H. and Van Way, J. (2016). Assessing levels of endangerment in the *Catalogue of Endangered Languages* (ELCat) using the Language Endangerment Index (LEI). *Language in Society*, 45:271–292.

Pérez Báez, G., Vogel, R., and Patolo, U. (2019). Global Survey of Revitalization Efforts: A mixed methods approach to understanding language revitalization practices. *Language Documentation and Conservation*, 13(1):446–513.

Thieberger, N. (2017). *LD&C* possibilities for the next decade. *Language Documentation and Conservation*, 11:1–4.

Proceedings, pages 112–115
Paris, UNESCO Headquarters, 5-6 December, 2019.

Importance of Frameworks in Language Technology - Case of Arabic

Karim Bouzoubaa
Mohammadia School of Engineers, Mohammed V University of Rabat, Morocco
karim.bouzoubaa@emi.ac.ma

Abstract

Arabic Language Technology has known a significant progress during the last years. As a result, several tools, resources and applications have been developed such as tokenizers, Part Of Speech taggers, morphological analyzers, syntactic parsers, etc. However, most of these tools are heterogeneous and can hardly be reused in the context of other projects without modifying their source code. This problem is known to be common to all rich-resourced languages, that is why some advanced frameworks have emerged such as GATE and UIMA. These frameworks have significantly changed the way Language Technology applications are designed and developed. They provide homogenous structures for applications, better reusability and faster deployment. In this paper, (i) we present a comparative study of frameworks in order to specify which ones can suitably deal with Arabic; and (ii) report on best practises to be applied for low-resourced langages

Keywords: Language Technology, Natural Language Processing, Rich and Low resourced languages, Arabic, Framework

Résumé

عرفت تقنية اللغة العربية تقدما ملحوظا خلال السنوات الماضية. ونتيجة لذلك، تم تطوير العديد من الأدوات والموارد والتطبيقات مثل مجزء الكلام، والمحللات الصرفية، والمحللات النحوية، وما إلى ذلك.

ومع ذلك، فإن معظم هذه الأدوات غير متجانسة ولا يمكن إعادة استخدامها في سياق مشاريع أخرى دون تعديل المصادر الخاصة .

من المعروف أن هذه المشكلة شائعة في جميع اللغات الغنية بالموارد، ولهذا السبب ظهرت بعض المنصات المتقدمة مثل GATE و UIMA.

لقد غيرت هذه المنصات بشكل كبير طريقة تصميم وتطوير تطبيقات تقنية اللغة. إنها توفر هياكل متجانسة للتطبيقات، وإعادة استخدام أفضل ونشر أسرع. في هذه الورقة: (1) نقدم دراسة مقارنة من أجل تحديد المنصات التي يمكن أن تتعامل بشكل مناسب مع اللغة العربية؛ و (2) نذكر قائمة أفضل الممارسات التي يجب تطبيقها على اللغات منخفضة الموارد

1. Introduction

One of the main issues to consider when developing any natural language processing system is choosing the most appropriate tool. For the case of Arabic language, many interesting development tools already exist: morphological analyzers to define the structure of words, stemmers to group words that have the same root or stem, syntactic parsers that determines the structures of sentences, and so on.

Most of the time, when it comes to different levels of a language (morphology, syntax, semantics and pragmatics) and because each level has its own specificity, many tools are needed at the same time. This usually leads to problems such as: integration of different technologies, more difficult maintenance of the system, a larger code in terms of the number of lines, a tedious search for appropriate solutions, etc. Thus, to avoid such problems, it is recommeded to have a single integrated framework allowing researchers to develop the different aspects of the language, and which proposes:

- Basic modules of automatic processing of the Arabic language such as morphological, syntactic and semantic analysis
- Free resources (dictionaries, corpora, lexical database, etc.) and modules for comparison and evaluation
- Applications for advanced processing such as sentiment analysis, fraud detection, plagiarism, and so on

In our view, frameworks represent an efficient way for standardization, optimization of efforts, collaboration and acceleration of developments in the field of NLP. Since our experience has mainly focused on the Arabic language, we briefly show how the Software Architecture for ARabic (SAFAR) framework can handle different levels of Arabic language.

This article is not a thorough presentation about SAFAR. The interested reader can refer to many aspects of the framework from (Jaafar et al. 2018, Jaafar and Bouzoubaa 2018, Namly et al. 2016). Herein, we aim reporting on two main issues (i) understand the importance of providing a framework for any language to boost the development of language technology in that language (ii) understand how to do it by considering the Arabic as the language of study and by benchmarking all existing frameworks and their integration of tools and resources.

The rest of this article is as follows. Section 2 describes the related works concerning framewords. Section 3 is dedicated to benchmarking these frameworks. Section 4 discusses main findings from the benchmarking. Finally, in the last section, we conclude the paper.

2. Related Works

In this section we present some known and commonly used NLP architectures that support partially/entirely Arabic, namely: GATE, UIMA, LIMA, Ling-Pipe, OpenNLP, NLTK, NooJ, ATKS, AraNLP, MADAMIRA and SAFAR. Since we want to focus mainly on Arabic language, we break these architectures into two main categories: (1) Independent language NLP architectures that support many NLP tasks and handle many languages including Arabic, (2) ANLP architectures that are specifically designed for processing Arabic.

2.1 Independent language NLP Architectures

We give general presentations of some independent language NLP architectures without focusing on Arabic language. These general presentations aim to introduce each architecture and get a clear idea about its structure and functionalities. This way, we will focus later on benchmarking these architectures regarding to their support of Arabic.

GATE[1] (Cunningham et al. 2011) is an infrastructure for the development and deployment of components for the processing of natural language. Developed since 1995 at the University of Sheffield, it is widely used for text mining and information extraction. GATE offers an architecture, a Java framework (including many modules) and an integrated development environment. However, it has some limitations: GATE components are too abstract, and do not offer a specification in terms of API and output for Arabic NLP components.

NOOJ[2] (Silberztein et al. 2012) is a language development environment for building, testing and maintaining formalized descriptions with wide coverage of natural languages (in the form of dictionaries and electronic grammars), and developing language processing applications. However, it adopts a unique formalism, based on automata, and is based on a pipeline architecture to form complex processes.

UIMA[3] (Ferrucci and Lally 2004) is a software architecture for the development and deployment of unstructured information analysis tools. Its purpose is to describe the processing steps of a text, image or video document in order to automatically extract structured information. The very general aim of this environment makes it a relatively low-level abstract architecture that does not offer, as such, any automatic language processing analysis module that can be used immediately. The implementation of processes for a given task thus remains the responsibility of the designer, who must be provided with analysis components developed by himself or by third parties.

OpenNLP[4] (Ingersoll et al. 2013) is a Java machine learning toolkit for the processing of natural language text. The main goal of the OpenNLP project is to create a mature toolkit for the most common NLP tasks: tokenization, sentence detection, part-of-speech (POS) tagging, named entity recognition, parsing, chunking, and coreference resolution. An additional goal is to provide a large number of pre-built model files for the aforementioned tasks.

NLTK[5] (Bird et al. 2008) is a platform for natural language processing and text analytics. The NLTK has originally been designed to support teaching in NLP and closely related areas. However, the NLTK has not only been used successfully for teaching, but also for prototyping and building Python programs to work with natural language data.

LingPipe[6], is a Natural Language Processing tool kit developed in Java by Alias-I[11] company. It is designed to be effective, extensible, reusable and robust. LingPipe performs tasks such as tokenization, Part-Of-Speech tagging, named entity recognition, clustering, database text mining etc.

LIMA[7] (Besançon et al. 2010), a multilingual framework for linguistic analysis and linguistic resources development and evaluation, is an NLP framework developed by the LVIC laboratory of CEA LIST.

2.2 Arabic NLP Architectures

We introduce Arabic NLP architectures that are intended to process only Arabic. We present their internal structures, components and functionalities. It should be noted that we have found only few number of such architectures for Arabic.

ATKS[8], Arabic Toolkit Service, is a set of NLP components targeting Arabic language that can be exploited as web services. It was developed by Microsoft within the Advanced Technology Lab in Cairo and consists of eight ANLP components: a morphological analyzer (SARF), a spell-checker, an auto corrector, a diacritizer, a named entity recognizer (NER), a colloquial to Arabic converter, a parser and a part-of-speech (POS) tagger. It should be noted also that these components are integrated into several Microsoft products and services such as Windows, Office and Bing.

MADAMIRA[9] (Pasha et al. 2014) is a toolkit for morphological analysis and disambiguation of Arabic and its dialects [5]. This toolkit provides seven ANLP processing tasks: Tokenization, morphological disambiguation for full range of morphological features, Part-of-Speech tagging, lemmatization, diacritization, named entity recognition and base phrase chunking. It should be noted that MADAMIRA is considered in our point of view as a fixed pipeline rather than a flexible

[1] https://gate.ac.uk/
[2] http://www.nooj-association.org/
[3] https://uima.apache.org/
[4] https://opennlp.apache.org/
[5] http://www.nltk.org
[6] http://www.alias-i.com/lingpipe/
[7] https://github.com/aymara/lima
[8] https://www.microsoft.com/en-us/research/project/arabic-toolkit-service-atks/
[9] https://camel.abudhabi.nyu.edu/madamira/

toolkit that separates components from each other and give at the same time the possibility to create other types of pipelines.

AraNLP[10] (Althobaiti et al. 2014) is a Java-based architecture that contains several Arabic text basic tools [4]. AraNLP attempts to bring together most of the vital Arabic text preprocessing tools into one single structure that can be accessed easily by end users. So far, AraNLP includes a sentence detector, tokenizer, light stemmer, root-based stemmer, part-of-speech tagger (POS-tagger), word segmenter, normalizer, and a punctuation and diacritic remover. Users can build customized pipelines by piping output from one tool into the next. For example, a customized pipeline can forward output from the sentence detector to the word segmenter into the POS tagger.

SAFAR[11] is a Java-based framework dedicated to Arabic. It brings together all layers of ANLP: resources, pre-processing, morphology, syntax and semantics. The general idea behind SAFAR is to gather, within a single homogeneous architecture, the available set of Arabic tools that are already developed, and develop new ones if necessary. Application builders can then realize many benefits by reusing components and avoid problems of interoperability as long as all components within the framework share the same architecture. All tools within SAFAR are standardized according to several Java interfaces. Therefore, users can add new implementations of any family tools just by implementing the appropriate interface. They could also easily create customizable pipelines where the output of one tool is the input of another. SAFAR outputs can be either memory objects or output files. So far, SAFAR contains several tools and resources such as morphological analyzers, stemmers, parsers, utilities, etc.

3. Benchmarking NLP Architectures

For this benchmark, we have selected several criteria that we have grouped into four features: 1) Arabic integrated tools, 2) Arabic integrated resources, 3) flexibility of exploitation and 4) maintenance and support. It should be noted that this benchmark concerns the Arabic side within each architecture and not all its aspects. For example, UIMA has too many published articles, but we are interested only on those that concern Arabic, the same goes for integrated tools and the other criteria. That is to say, tables below aim to present how these architectures are concerned by Arabic and how they handle it.

As it is shown in table 1, NLP architectures dedicated to Arabic exceed largely independent language architectures in terms of Arabic integrated tools. This is obvious because ANLP architectures are intended to contain only Arabic tools unlike others. Indeed, the ANLP community is not encouraged to integrate its works within such language-independent architectures. This justifies the lack of Arabic language processing components within these architectures. SAFAR framework comes in the first place since it implements various tools within all its layers.

Unlike tools, resources come with fewer numbers. This is because they are time consuming when developing them comparing to some tools such as toknizers and light stemmers. Resources also require the cooperation of computer scientists as well as linguists, which complicates the task especially for huge resources. Table 1 shows that each of LIMA and NooJ provide only one resource which are respectively a lexicon and corpora. SAFAR is the only architecture that provides many Arabic resources. Some architectures provide no Arabic resource, while others provide resources (such as clitics, roots, etc.) but used in the context of their programs and it is up to the programmer to understand the workflow of the program and extract the corresponding resource.

Architecture	Tools	Resources	Flexibility	Maint.
UIMA	1	0	10	3
GATE	6	0	10	2
LIMA	1	1	7	2
Nooj	0	1	6	2
SAFAR	23	8	10	16
AraNLP	8	0	4	2
OpenNLP	0	0	5	1
NLTK	1	0	3	2
LingPipe	1	0	5	1
ATKS	8	0	4	1
MADAMIRA	7	0	6	2

Table 1: Architectures according to four features

The flexibility score is calculated from four different metrics which are (1) The number of possible data formats available to processes resources; (2) The possibility or not to extend the architecture and program pipelines; (3) The possibility to exploit it as an API and/or via web services; and (4) the portability or not of the architecture. UIMA, GATE and SAFAR are getting the highest score since they provide all the above mentioned features.

The score of maintenance is calculated by attributing one point for each architecture that has a release within the last five years, one point if it has documentation and one point for every Arabic published article. Table 5 shows also that most of the presented architectures are up to date regarding their releases. However, SAFAR is the only architecture that provides many published articles concerning Arabic. Each one of these articles addresses one or many aspects of processing Arabic within SAFAR. Other architectures do not provide any articles or provide few ones focusing on Arabic language as for LIMA, NooJ, AraNLP and MADAMIRA. Concerning the documentation, UIMA, GATE, SAFAR and NLTK have extensive ones. This can be very helpful to get started and be familiar with these architectures with a minimum effort from end users. Indeed, less documentation leads to more effort to discover how to manipulate it, and vice versa.

[10] https://sites.google.com/site/mahajalthobaiti/resources
[11] http://arabic.emi.ac.ma/safar/

4. Discussion

We can classify all architectures into four main categories namely zone “a”, “b”, “c” and “d” (figure 1). Zone “1” represents all architectures which are dedicated to specific processing including but not limited to Arabic. Zone “2” represents all advanced architectures used to handle complex processing either for Arabic or other languages. Zone “3” concerns all architectures which are dedicated specifically to Arabic.

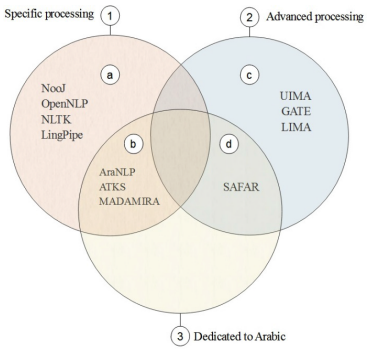


Figure 1: Architectures according to researchers needs

These three zones are general and can be intersected to produce new zones for more detailed and specific needs. For example, zone “a” is equal to zone “1” minus zone “b”, which represents all architectures for specific processing that are not dedicated to Arabic namely: NooJ, OpenNLP, NLTK and LingPipe. The same goes as well for zone “b”, which is the intersection of zone “1” and zone “3” and which represents all architectures for specific processing that are dedicated specifically to Arabic, namely: AraNLP, ATKS and MADAMIRA. In the other side, zone “c” is equal to zone “2” minus zone “d”, which represents all advanced architectures that are not dedicated to Arabic, namely UIMA, GATE and LIMA. Finally, zone “d” is the intersection of zone “2” and zone “3” and which represents architectures for advanced processing that are dedicated specifically to Arabic, namely SAFAR.

Hence, we conclude that all the presented architectures can be used according to researches needs and specific contexts. However we believe that frameworks are an important step towards standardization, the resolution of interoperability issues, reusability and integration of all development efforts in the field of language processing.

From the above study and comparison between frameworks for the specific case of Arabic, we learn:

- Rich-resource languages have plenty of frameworks (many others exist in addition to those presented in this paper) to consider when developing for language technology purposes
- For low-resourced language, either no frameworks exist such as the Amazigh language (local language spoken mainly in Maghreb region), or very few components (tools and resources) are integrated in known and large frameworks
- For low-resourced langagues, we recommend first of all to compile all existing and available components and then group them in a framework. This way, researchers would find a central point of development and would avoid develop redundant componants. As a consequence, this will boost language technology development in that langauge.

5. Conclusion

In this paper, we presented a comparative study of Natural Language Processing architectures that can handle Arabic language. Indeed these architectures represent for us a way to standardize the various aspects shared by Arabic processing tools in order to promote interoperability. For this, we have highlighted several architectures among which we cite UIMA, GATE, AraNLP and SAFAR.

6. References

M. Althobaiti, U. Kruschwitz and M. Poesio, "AraNLP: a Java-Based Library for the Processing of Arabic Text," in Proceedings of the 9th Language Resources and Evaluation Conference (LREC), Reykjavik, 2014.

R. Besançon, G. De Chalendar, O. Ferret, F. Gara, O. Mesnard, M. Laïb and N. Semmar, "LIMA: A Multilingual Framework for Linguistic Analysis and Linguistic Resources Development and Evaluation," in Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, 2010.

S. Bird, E. Klein, E. Loper and J. Baldridge, "Multidisciplinary instruction with the natural language toolkit," in Proceedings of the Third Workshop on Issues in Teaching Computational Linguistics, 2008.

H. Cunningham, D. Maynard, K. Bontcheva and others, Text processing with gate, *Gateway Press* CA, 2011.

D. Ferrucci and A. Lally, "Building an example application with the unstructured information management architecture," *IBM Systems Journal*, vol. 43, no. 3, pp. 455-475, 2004.

G. S. Ingersoll, T. S. Morton and A. L. Farris, Taming text: how to find, organize, and manipulate it, *Manning Publications* Co., 2013.

Y. Jaafar, M. Nasri, K. Bouzoubaa, "Semantic Analysis of Arabic Texts within SAFAR Framework ", 5th International IEEE Congress on Information Science and Technology CIST'18, Marrakech, Morocco, October 2018

Y. Jaafar, K. Bouzoubaa, "A Survey and Comparative Study of Arabic NLP Architectures". In: Shaalan K., Hassanien A., Tolba F. (eds) *Intelligent Natural Language Processing: Trends and Applications. Studies in Computational Intelligence*, vol 740. Springer, Cham, 2018

D. Namli, Y. Regragui, K. Bouzoubaa, "Interoperable Arabic language resources building and exploitation in SAFAR platform", In Proceeding of the 13th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA 2016), Agadir, Morocco, November 2016.

A. Pasha, M. Al-Badrashiny, M. Diab, A. El Kholy, R. Eskander, N. Habash, M. Pooleery, O. Rambow and R. M. Roth, "MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic," in LREC'14, Reykjavik, 2014.

M. Silberztein, T. Varadi and M. Tadic, "Open source multi-platform NooJ for NLP," in COLING (Demos), Mumbai, 2012.

Proceedings, pages 116–119
Paris, UNESCO Headquarters, 5-6 December, 2019.

Promoting Language Technology for Endangered Languages with Shared Tasks

Gina-Anne Levow
University of Washington
Seattle, WA USA
levow@uw.edu

Abstract

Although recent years have seen dramatic improvements in speech and language technologies, such systems are only available for the few hundred highest resource languages. These systems' reliance on large annotated datasets has limited their impact on the thousands of low-resource and endangered languages which could otherwise benefit. At the same time, it is estimated (Woodbury, 2019) that 50-80% of the world's languages are at risk of disappearing by 2100. Shared tasks have helped to drive rapid development of language technologies by providing shared, standardized datasets, evaluation metrics, and venues to communicate approaches and results. The current project aims to develop shared tasks and systems targeting language technologies that can accelerate documentation and facilitate revitalization of endangered languages, while advancing the state of the art.

Keywords: Shared Tasks, speech processing, speaker recognition, language identification

1. Introduction[1]

Speech and language processing techniques have made dramatic strides in the recent years. Human "parity", system performance at the level of people performing the same tasks, has been claimed for high profile tasks, such as automatic speech-to-text transcription (Stolcke and Droppo, 2017; Xiong et al., 2017) and machine translation (Hassan et al., 2018). However, these highly impressive results have typically been demonstrated on high-resuource languages such as English and the Chinese-English pair in the examples above. Furthermore, web-based tools for these applications are available for only 100-200 languages. Unfortunately, this situation is not surprising as the success of these systems relies not only on algorithmic advances, but also crucially on large-scale language resources, on the order of thousands or even tens of thousands of hours of transcribed speech or tens of millions of lines translated text.

As a result, it remains difficult to bring these new technologies to bear to benefit lower resource or endangered languages. At the same time it is feared that of the more than 6,000 languages spoken around the world, 50-80% could be lost by 2100(Woodbury, 2019). Language technologies have the potential to dramatically accelerate and facilitate efforts in language documentation and revitalization, if they can be effectively harnessed.

2. Approach: Shared Tasks

To bridge this gap between speech and language technologies and the needs of endangered language researchers and speaker communities, we plan to leverage the framework of Shared Task Evaluation Campaigns (STECs or Shared Tasks). STECs have been powerful drivers of speech and language technologies (Belz and Kilgarriff, 2006) and have contributed to the development of systems ranging from spoken dialog systems (Dahl et al., 1994) to document retrieval (Voorhees and Harman, 2005) to information extraction(Grishman and Sundheim, 1996). STECs provide standard datasets for training and testing systems, standard evaluation metrics, and venues for sharing results and techniques. As such, these tasks allow access to valuable data needed for system creation, direct comparisons across different methods, and transmisson of successful strategies. They also allow the community and organizers to focus research on tasks of interest, while pooling the costs of resource development.

Our planned STECs will focus research attention on tasks which will benefit endangered language researchers and speaker communities and will leverage growing archives of endangered language data. This setting will enable researchers in speech technology to assess their systems on a broader and more diverse range of languages than are typically employed (Bender, 2011). The tasks will also provide a more naturalistic setting in which to evaluate models for low-resource language processing, in contrast to simulations of low-resource settings by subsetting high resource language data.

3. Background & Design Principles

To help design tasks that would address crucial needs among researchers in endangered languages and speaker communities while challenging the state-of-the-art in language technology, a National Science Foundation-funded workshop, EL-STEC: Shared Tasks with Endangered Language Data, was held in September 2016, bringing together representatives of these different communitites. The discussions were driven by the identification of key pain points in the workflow of those striving to understand and document endangered languages, as well as capabilities desired by speaker communities. This process helped to define a suite of candidate tasks. In addition, it led us to articulate design principles that guide the structure of these and subsequent STECs. These criteria are described below.

[1]This paper summarizes and updates an original publication in (Levow et al., 2017).

Realism Our tasks should reflect the needs and usage environments of future users. We will also encourage participants to leverage any available resources for these tasks, rather than artificially restricting these resources as is very common in other shared task settings. Such sources range from linguistic repositories, such as ODIN (Lewis and Xia, 2010) or WALS (Haspelmath et al., 2008), to models derived from high resource languages that could be adapted to new tasks, such as pre-built Universal Background Models (Hasan and Hansen, 2011) or X-vector (Snyder et al., 2018) models for speaker identification. The structure of existing archive data will intrinsically impose a range of challenges, including limited data size, varied recording or collection conditions, differences in speakers or genres, and so on.

Typological diversity To truly focus on language technologies with broad effectiveness and applicability, our tasks will involve multiple typologically distinct languages in training and testing. In addition, evaluation will include previously unseen "surprise" languages from additional language families which will explicitly test new systems' portability to new languages.

Accessibility of shared tasks We hope to encourage broad participation in these tasks by lowering barriers to entry. In particular, baseline systems will be provided as part of the shared tasks to all participants, in addition to data and evaluation software. These baseline systems allow the organizers to validate the data and task design, as well as to provide a reference level of effectiveness. Since the baselines will be released publicly, they can also serve as starting points for participating teams with fewer resources, e.g. students, to build on to develop their own submissions. Finally, following the model of the Speech Recognition Virtual Kitchen (SRVK) project (Plummer et al., 2014), we will encapsulate these baselines and all needed software for the shared tasks in virtual machines to facilitate cross-platform development.

Accessibility of resulting systems It is crucial that the technologies developed under these shared tasks become available to benefit new user populations, including endangered language researchers and speaker communities. A key requirement is thus that participating teams provide detailed descriptions of their systems to enable replication and reimplementation. We also encourage the creation of systems based on free or open-source software.

Extensibility These initial tasks will establish a first set of multilingual resources for endangered languages as well as baseline performance against which to measure future progress. In addition, the tools developed for dataset preparation will provide a template for extension to additional new languages in future years.

Nuanced evaluation Multiple metrics can be incorporated in evaluation to better assess the strengths and weaknesses of different approaches or different facets of the tasks, rather than focusing on a single metric and leaderboard rank.

The EL-STEC workshop working groups defined three tasks which aimed to embody the above principles: one focused on speech processing, one on morpho-syntactic analysis and glossing through first-pass creation of interlinear glossed text, and one on orthographic normalization for endangered languages. Below, we describe the first set of tasks that we plan to field in upcoming challenges.

4. "Grandma's Hatbox": Speech Processing for endangered languages

A core challenge in documenting endangered languages is that although many hours of recordings between a linguist and their consultant(s) may be collected, those valuable recordings must still pass through a lengthy and time-consuming series of steps including segmentation, transcription, alignment, and glossing. Due to the time and effort required, each stage of this process yields less and less material analyzed in greater and greater detail, in a sort of funneling. We define a set of tasks that could help to increase the throughput of this process by automating key steps of speech processing, thereby freeing those working with endangered languages to focus their expertise on important areas of analysis, while making more material available to archives, and providing richer metadata to support easier access for speaker communities.

The "Grandma's Hatbox" task cascade envisions the process required to prepare and archive a new collection of speech recordings. The name evokes a scenario where a trove of recordings and field notes is discovered left behind in a box or donated by a former researcher. The subtasks involved are as follows:

- segmentation of the recordings by language, considering both automatic identification of known high and low resource languages and clustering of languages not known in advance,
- segmentation of the recordings by speaker, considering both automatic identification of known speakers and clustering of speakers not known in advance,
- automatic identification of the genre of the recording, to be drawn from a small fixed inventory, including narrative, conversation, elicitation, and ritual, and
- finally, automatic alignment of partial transcriptions to recorded audio.

Participating teams would be able to choose to work on any subsets of these tasks. Each step above can feed into a subsequent processing step to create an enriched audio archive. The resulting metadata can be provided as a template to augment an archive's database or converted to a standard format viewable in an interactive interface such as ELAN (Brugman and Russel, 2004). A graphical depiction of the automatic processing and enrichment appears in Figure 1.

5. Challenging the State of the Art in Speech Processing

The tasks outlined above simultaneously address key pain points in the work process of field linguists and others who work with endangered languages and push the state of the art in spoken language processing. By working with endangered language data, these tasks pose novel challenges and

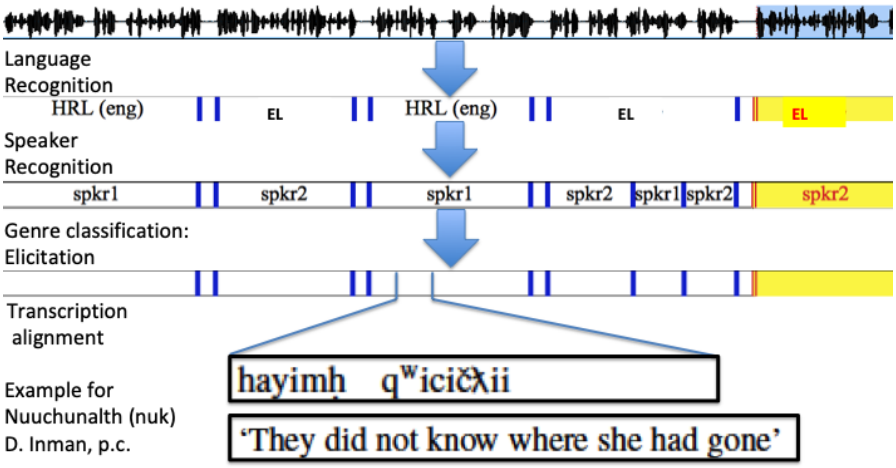


Figure 1: Speech processing cascade applied to audio and transcriptions.

present important opportunities. Broadly, they facilitate assessment of existing methods on new languages, many of which are typologically distinct from those on which most speech tools are trained and tested. Furthermore, the rise of neural network models has demonstrated effective techniques for exploiting large-scale data resources; however, there is substantial interest in low-resource techniques, such as those using unsupervised or semi-supervised learning. The fundamental resource constraints present in work with endangered language data will encourage exploitation and development of such methods.

Each of the steps of the speech processing cascade builds on existing technologies, spanning language identification, speaker diarization, speaker recognition, and automatic alignment. Although there are shared regimes in some of these areas, such as the NIST speaker (NIST, 2019) and language (NIST, 2017) recognition evaluations, the character of endangered language data poses new and important challenges beyond those typically addressed. This new data often involves multiple speakers, speaking in multiple languages or dialect varieties, often with short turns and possibly with fine-grained code-mixing. Rather than the broadcast or telephone conversational speech typically found in speech research corpora, recordings of endangered language data involve a range of genres and diverse, possibly noisy, recording conditions. These factors are often listed when discussing the limitations of existing tools. By creating tasks which drive development of systems to address these challenges, our STECs will advance these technologies.

6. Potential Benefits to Speaker Communities

In addition to aiding researchers working with endangered languages, these shared tasks hold the potential to also benefit speaker communities and aid in revitalization efforts. The speech processing tools presented above can facilitate access to new or existing recordings within communities. Speech, and video, recordings are notoriously slow and difficult to search or browse. By identifying the languages, speakers, and genres of recordings, the tools can enable some search and navigation within these valuable materials. In conjunction with speech alignment, content-based search and browsing could also be supported.

7. Conclusions & Future Work

We believe that Shared Task Evaluation Campaigns designed around endangered language data have the potential to benefit field linguists, endangered language researchers, language archives, and speaker communities while driving improvements in language technology. We anticipate that these enhancements will yield technologies that have applicability to a more diverse range of languages, both in terms of typology and in terms of availability of linguistic resources. We are currently preparing the datasets and software to launch the first iteration of our STECs in the coming year. We look forward to engaging with new partners and potential task participants.

8. Acknowledgements

This work has been supported by NSF #: 1500157 and NSF #: 1760475. Any opinions expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Great thanks to the members of the STREANLInED team at the University of Washington, especially my co-PI Emily M. Bender, Isaac Manrique, Jiani Chen, and Harita Kannan. We are also grateful for the contributions of Shobhana Chelliah, Joshua Crowgey, Dan Garrette, Jeff Good, Sharon Hargus, Mark Hasegawa-Johnson, Kristen Howell, Russ Hugo, David Inman, Jeremy Kahn, Lori Levin, Patrick Littell, Michael Maxwell, Alexis Palmer, Michael Tjalve, Laura Welcher, and Fei Xia during the EL-STEC workshop.

9. Bibliographical References

Belz, A. and Kilgarriff, A. (2006). Shared-task evaluations in HLT: Lessons for NLG. In *Proceedings of the Fourth International Natural Language Generation Conference*,

pages 133–135, Sydney, Australia. Association for Computational Linguistics.

Bender, E. M. (2011). On achieving and evaluating language independence in NLP. *Linguistic Issues in Language Technology*, 6:1–26.

Brugman, H. and Russel, A. (2004). Annotating multimedia/ multi-modal resources with ELAN. In *Proceedings of LREC 2004, Fourth International Conference on Language Resources and Evaluation*.

Dahl, D. A., Bates, M., Brown, M., Fisher, W., Hunicke-Smith, K., Pallett, D., Rudnicky, A., and Shriberg, E. (1994). Expanding the scope of the ATIS task: the ATIS-3 corpus. In *Proceedings of the ARPA Human Language Technology Workshop*, pages 43–48. Morgan Kaufmann.

Grishman, R. and Sundheim, B. (1996). Message understanding conference - 6: A brief history. In *Proceedings of Coling 1996*, pages 466–471.

Hasan, T. and Hansen, J. H. L. (2011). A study on universal background model training in speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):1890–1899.

Martin Haspelmath, et al., editors. (2008). *The World Atlas of Language Structures Online*. Max Planck Digital Library, Munich. http://wals.info.

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M., Liu, S., Liu, T.-Y., Luo, R., Menezes, A., Qin, T., Seide, F., Tan, X., Tian, F., Wu, L., Wu, S., Xia, Y., Zhang, D., Zhang, Z., and Zhou, M. (2018). Achieving human parity on automatic Chinese to English news translation.

Levow, G.-A., Bender, E. M., Littell, P., Howell, K., Chelliah, S., Crowgey, J., Garrette, D., Good, J., Hargus, S., Inman, D., Maxwell, M., Tjalve, M., and Xia, F. (2017). STREAMLInED challenges: Aligning research interests with shared tasks. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 39–47, Honolulu, March. Association for Computational Linguistics.

Lewis, W. D. and Xia, F. (2010). Developing ODIN: A multilingual repository of annotated language data for hundreds of the world's languages. *Journal of Literary and Linguistic Computing*, 25:303–319.

NIST. (2017). 2017 Language Recognition Evaluation Plan. `https://www.nist.gov/system/files/documents/2017/09/29/lre17_eval_plan-2017-09-29_v1.pdf`. Downloaded November 21, 2019.

NIST. (2019). 2019 NIST Speaker Recognition Evaluation Plan. `https://www.nist.gov/system/files/documents/2019/07/22/2019_nist_speaker_recognition_challenge_v8.pdf`. Downloaded November 21, 2019.

Plummer, A., Riebling, E., Kumar, A., Metze, F., Fosler-Lussier, E., and Bates, R. (2014). The Speech Recognition Virtual Kitchen: Launch party. In *Proceedings of Interspeech 2014*.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S. (2018). X-vectors: Robust DNN embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Stolcke, A. and Droppo, J. (2017). Comparing human and machine errors in conversational speech transcription. In *Proc. Interspeech 2017*, pages 137–141.

Ellen M. Voorhees et al., editors. (2005). *TREC: Experiment and Evaluation in Information Retrieval*. Digital libraries and electronic publishing series. The MIT Press, Cambridge, MA.

Woodbury, A. C. (2019). What is an endangered language? Accessed November 2019.

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M. L., Stolcke, A., Yu, D., and Zweig, G. (2017). Toward human parity in conversational speech recognition. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 25(12):2410–2423, December.

Proceedings, pages 120–122
Paris, UNESCO Headquarters, 5-6 December, 2019.

Towards a Global Lexicographic Infrastructure

Simon Krek[1], Thierry Declerck[2,3], John McCrae[4], Tanja Wissik[3]
[1]Jožef Stefan Institute, Slovenia
[2]DFKI GmbH, Multilinguality and Language Technology, Germany
[3]Austrian Centre for Digital Humanities at the Austrian Academy of Sciences, Austria
[4]Insight Centre for Data Analytics at the National University of Ireland Galway, Ireland
[1]simon.krek@guest.arnes.si
[2]declerck@dfki.de
[3]Tanja.Wissik@oeaw.ac.at
[4]john.mccrae@insight-centre.org

Abstract

In this paper we briefly describe the European project ELEXIS (European Lexicographic Infrastructure). ELEXIS aims to integrate, extend and harmonise national and regional efforts in the field of lexicography, both modern and historical, with the goal of creating a sustainable infrastructure which will enable efficient access to high quality lexical data in the digital age, and bridge the gap between more advanced and lesser-supported lexicographic resources. For this, ELEXIS makes use of or establish common standards and solutions for the development of lexicographic resources and develop strategies and tools for extracting, structuring and linking lexicographic resources. This paper is a kind of summary of a more technical description of ELEXIS included in the proceedings of the Globalex 2018 workshop.

Keywords: eLexicography, Standards, Infrastructure

Résumé

Dans cet article, nous décrivons brièvement le projet européen ELEXIS (infrastructure lexicographique européenne). ELEXIS a pour objectif d'intégrer, d'étendre et d'harmoniser les efforts nationaux et régionaux dans le domaine de la lexicographie, à la fois moderne et historique. Le but est de créer une infrastructure durable qui permettra un accès efficace à des données lexicales de haute qualité à l'ère numérique et de combler le fossé entre les ressources lexicographiques les plus avancées et celles beaucoup moins développées. Pour ce faire, ELEXIS utilise ou établit des normes et des solutions communes pour le développement de ressources lexicographiques et élabore des stratégies et des outils pour extraire, structurer et relier les ressources lexicographiques. Cet article est une sorte de résumé d'une presentation plus technique d' ELEXIS, qui est include dans les actes du workshop Globalex 2018.

1. Introduction

The field of lexicography has a long tradition of proposing as accurate as possible descriptions of languages. As stated in (Køhler Simonsen, 2017): "Lexicography is a four thousand-year-old discipline and dictionaries have been an integral part of commerce and human cultural history for centuries".

Since the 1980s, lexicographers have started to utilize computers and to apply computational methods. Online dictionaries are no longer only a reference work but are also seen as platforms for supporting advanced search facilities. This emerging field of e-lexicography, nevertheless, is still not clearly shaped, and methods and workflows not yet fully agreed on. Michael Rundell (2015) for example describes the current situation of e-lexicography as being in a transitional phase. A quotation of Robert Lew stating that "It seems that the web community, while enthusiastically embracing the novelty of online collaboration, propagates the traditional model of lexicographic description"[1].

In recent years, however, new developments have emerged in the field of e-lexicography, like the eLex conference se-ries[2], which started in 2009, the Globalex initiative[3], which was established at eLex 2015 or the European Network of e-Lexicography (ENeL) COST action[4], which in 2013 brought together for the first time a large number of lexicographers to discuss issues related to the emergence of new technologies. ENeL was set up to improve the access for the general public to scholarly dictionaries and make them more widely known to a larger audience. In the context of this network, a clear need emerged for a broader and more systematic exchange of expertise, for the establishment of common standards and solutions for the development and integration of lexicographical resources, and for broadening the scope of application of these high quality resources to a larger community, including the Semantic Web, Artificial Intelligence, Natural Language Processing and Digital Humanities. This is where ELEXIS comes into play.

2. ELEXIS

ELEXIS (European Lexicographic Infrastructure) is fostering cooperation and information exchange among lexicographical research communities. The infrastructure is a granted project under the H2020-INFRAIA-2016-2017 call, with the topic "Integrating Activities for Starting Communities" and started in February 2018[5]. ELEXIS is building on infrastructures defined in other projects and initiatives, especially CLARIN[6] and DARIAH[7], which allow language or Digital Humanities resources (both tools and data) to be shared.

A key goal of the ELEXIS project is thus to enable stakeholders to link their existing lexicographic resources, either as dictionaries or as standalone lexical descriptions encoded, and so to create a huge multilingual registry, a kind of "Matrix Dictionary" that connects lexicographic resources across common concepts.

2.1 A Matrix Dictionary for ELEXIS

A key goal of ELEXIS is the creation of a "Matrix Dictionary", that is formed of links created between lexicographic resources in different languages, domains and forms, independently if the language considered is high- or under-resourced. With this, ELEXIS is creating a universal repository of linked senses, meaning descriptions, etymological data, collocations, phraseology, translation equivalents, examples of usage and all other types of lexical information found in all types of existing lexicographic resources, monolingual, multilingual, modern, historical, etc.

In order to reach this goal, ELEXIS makes uses of strategies, tools and standards for extracting, structuring and linking the high-quality semantic data from lexicographic resources and make them available to the Linked (Open) Data[8] family.

Those processes are necessary, as current lexicographic resources, both modern and historical, have different levels of structure and are not equally suitable for applications in advanced NLP technologies, like Information Retrieval or Machine Translation, for which they should be disclosed to or from which they could benefit. The project works also on interlinking lexical content with other structured or unstructured data – corpora, multimodal resources, etc. – on any level of lexicographic description: semantic, syntactic, collocational, phraseological, etymological, translation equivalents, examples of usage, etc.

[1] Taken from taken from (Lew, 2014).

[2] See https://elex.link/

[3] See https://globalex.link/

[4] See https://www.elexicography.eu/

[5] See http://www.elex.is/

[6] See https://www.clarin.eu/.

[7] See https://www.dariah.eu/.

[8] See https://www.lod-cloud.net/ and also https://linguistic-lod.org/ for the subset of the LOD dealing with linguistic data.

ELEXIS conversion and alignment tools will provide users of the infrastructure with the possibility to harmonise and convert their lexicographic resources to a uniform data format that allows their seamless integration in Linked Open Data or in other repositories.

2.2 The virtuous Cycle of e-Lexicography in ELEXIS

ELEXIS implements a cyclic approach to the building and linking of lexicographic resources. We use the term "virtuous circle" for this, as it characterizes an integrative approach to a spiralling development of lexicographic data on the basis of a cross-disciplinary exchange of knowledge and the incremental contributions of the different methods and technologies to be involved. Figure 1 is given a graphical representation of this cyclic development.

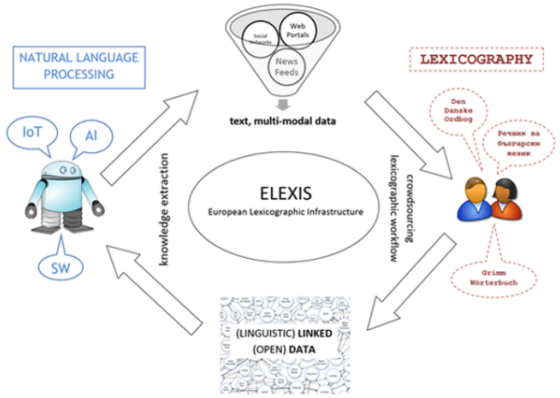


Figure 1: The virtuous cycle of e-lexicography

The existence of common data models and standards that are produced bottom-up from within the lexicographic community fostered by ELEXIS is a necessary condition for the successful development of the whole platform. Standards are developed and tested during the project on the data provided by the lexicographic partners and implemented in the newly developed service.

3. Relevance for Less-Resourced Languages

ELEXIS supports novel lexicography by providing lexicographers with tools and methods that help them create new resources. Using machine learning, data mining and information extraction techniques proto-dictionary content will be produced in an automated way. The automatically extracted data can then be used as a starting point for further processing either in a more traditional lexicographic workflow or through crowdsourcing platforms, making it easier to create new resources. This novel approach can be applied to any language for which there is data available on the web. This is particularly important for under-resourced languages with out-dated or non-existent language descriptions, enabling researchers and the general public to learn about semantic, grammatical or other aspects of lexicographic description benefiting from the technology derived from language communities with advanced lexicographic descriptions.

4. Summary

ELEXIS is aiming at the following points:

- foster cooperation and knowledge exchange between different research communities in lexicography in order to bridge the gap between lesser-resourced languages and those with advanced e-lexicographic experience;
- establish common standards and solutions for the development of lexicographic resources;
- develop strategies, tools and standards for extracting, structuring and linking of lexicographic resources;
- enable access to standards, methods, lexicographic data and tools for scientific communities, industries and other stakeholders;
- promote an open access culture in lexicography, in line with the European Commission Recommendation on access to and preservation of scientific information.

ELEXIS is based on the conviction that lowering the barrier for retrieving and analysing multilingual lexicographic data across Europe – and beyond cannot be accomplished in the long term without lowering the barrier for providing lexicographic data to research infrastructures.

5. Acknowledgements

The ELEXIS project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

6. Bibliographical References

Kilgarriff, A. (2000). Business models for dictionaries and nlp. International Journal of Lexicography, 13(2):107-118.

Køhler Simonsen, H. (2017). Lexicography: What is thebusiness model? In Iztok Kosem, et al., editors, Electronic Lexicography in the 21st Century, pages 395–415. Lexica Computing CZ s.r.o.

Lew, R. (2014). User-generated content (ugc) in online English dictionaries. OPAL - Online publizierte Arbeiten zur Linguistik, 4:8–16.

Rundell, M. (2015). From print to digital: Implications for dictionary policy and lexicographic conventions. Lexikos, 25(1).

Proceedings, pages 123–126
Paris, UNESCO Headquarters, 5-6 December, 2019.

Language Technologies for Less-Resourced-Languages LRL Workshop Series at the Language and Technology Conferences (LTC) from 2009 to 2019

Zygmunt Vetulani[1], Khalid Choukri[2], Joseph Mariani[3] and Patrick Paroubek[4]

Adam Mickiewicz University in Poznań[1], ELRA-ELDA[2], LIMSI/CNRS[3,4]
ul. Uniwersytetu Poznańskiego 4, 61-614 Poznań, Poland[1], 9 Rue des Cordelières, 75013 Paris, France[2], Campus universitaire bât 507, Rue du Belvedère, F - 91405 Orsay cedex[3,4]
vetulani@amu.edu.pl, choukri@elda.org, joseph.mariani@limsi.fr, pap@limsi.fr

Abstract

When LTC started in Poznań, Poland, as LT Awareness Days (1995), Polish was still a "less-resourced-language". This report presents the LRL Workshop Series organized since 2009 as integral part of LTC. We present as raison d'être of LRL to contribute to a "roadmap towards supplying LR and LT for all languages". We go one by one through all LRLs (until 2019) to present themes suggested by organizers, affiliation countries of the authors, as well as the concerned languages. We note positive phenomena such as appearance of countries and languages so far very rare at the international LT conferences.

Keywords: language resources, less-resourced-languages, language technologies, LTC

Résumé

La première conférence LTC, tenue à Poznań (Pologne) en 1995, portait le nom "L&T Awareness Days", et à cette époque, le polonais faisait partie des langues peu dotées. Nous présentons ici la série des ateliers sur les langues peu dotées (LRL Workshop Series) qui font partie intégrale de la conférence LTC depuis 2009. LRL a pour sa raison d'être d'apporter une contribution à la feuile de route pour doter toutes les langues de resources et technologies du langage. Nous parcourons un à un les ateliers LRLs jusqu'en 2019 pour rappeler les thèmes proposés par les organisateurs, les pays d'affiliation des auteurs et les langues concernées. Nous notons avec satisfaction l' émergence à LTC de pays et de langues rarement vues à l'international.

1. Introduction

The Language and Technology Conferences (LTC) started in Poznań, Poland, in 1995. The first event of this series were Language and Technology Awareness Days co-organized by Adam Mickiewicz University, Poznań and the European Commission – DG XIII. At those days the Polish language was clearly a "less-resourced language" with, however, a solid traditional linguistic description based on strong logical and mathematical background. Similar situation was in other central and East European countries. During the next 15 years the golden age of Language Technologies continued and resulted in reduction of the initial technological gap between these countries. This was due to the intensive international collaboration within joint projects and individual mobility of experts and researchers. The series of LTC events, supported by ELRA/Elda, FlareNet and Meta-Net, were part of the global effort to foster integration of the LT community in Europe and abroad, for all languages

Since 2005 until now the LTCs were organized every two years as an international forum, open for both academia and language industry – these two communities together contributing to the development and dissemination of language technologies. The first global-scale event of 2005 mobilized over 100 contributors from Europe, Asia and North America. It was dedicated to the memory of Maurice Gross and Antonio Zampolli – two visionary personalities among the first to understand the necessity to bring together the two communities to favor the emerging language industries.[1]

But with dynamic development of language technologies and language industry, as well as amplification of globalization trends, the new threats of technological exclusion become more present in a world of fierce economic competition because of the world population increase and dwindling natural resources.

2. LRL Workshops

2.1 Kick-of

In 2009, we decided to attract more attention of the LT community to the case of the technologically "under-resourced" countries and menaced by technological, and successively, cultural exclusion. This idea gave birth to the Less-Resourced Languages Workshops (LRL) as being an integral part[2] of the LTC conferences. The first of them, co-chaired by Khalid Choukri, Joseph Mariani and Zygmunt Vetulani, was entitled "Getting Less-Resourced Languages on Board! " The rationale of this workshop was the following:

[1] We adhere to this idea with the full title of LTC which is: "Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics".

[2] "Integral part" means that LRL is open for all participants registered directly to all other workshops and tracks of the LTC and *vice versa*. Also acceptance procedures for LRL submissions are strictly the same as for all LTC attenders (see www.ltc.amu.edu.pl).

"Language Technologies (LT) provide an essential support to the challenge of Multilingualism. In order to develop them, it is necessary to have access to Language Resources (LR) and to assess LT performances. To this regard, the situation is very different across the different language. Little or sparse data exist for languages in countries or regions where limited efforts have been devoted to such issues in the past, also known as Less-Resourced Languages (LRL). The workshop aims at reporting the needs, at presenting achievements and at proposing solutions for the future, both in terms of LR and of LT evaluation, especially in the European, Euro-Mediterranean and regional frameworks. This will allow to identify the factors that have an impact on a potential and shared roadmap towards supplying LR and LT for all languages." (www.ltc.amu.edu.pl/2009/).

Ten years later this rationale still remains valid.

This half-a-day workshop, open to the whole LTC audience, was organized into three parts: invited talks, presentation of technical papers, and panel discussion to summarize the workshop. The first one was given by Briony Williams (Bangor Univ., Wales and ISCA) who presented a talk "Less-Resourced Languages and Language Resources: Lessons learned from the Celtic languages of Great Britain and Ireland." The second invited lecturer was Aleksandra Wesolowska (EC, DG – Information society and media) who addressed an essential question to the concerned audience: "New start for European language technology. Are you ready?". The second part consisted in presentation of 8 technical papers (32 authors) presented by teams from Bulgaria, France, Germany, Ghana, Greece, Norway, Slovak Republic, Spain, (with contributions on Amharic, Basque, Bulgarian, Catalan, Ga, Galician, Luxembourgish, Romani). The panel discussion with the motto ***"Linguists need technologists, technologists need linguists, societies and cultures need both to survive"*** concluded the workshop with a number of observations that were collected in the final workshop report[3]:

- a strong political will to consider the language dimension and sufficient funds are necessary,
- this must go with the awareness that Language Technologies and Language Resources are essential to the development of society,
- there should be specialists in the processing of any given language, reaching a critical mass, and young researchers should be trained.
- an infrastructure must exist, including:
  - a writing system/a transcription code/an agreed orthography,
  - Language Resources (sufficient in quantity and quality),
  - tools (especially language independent ones, if possible as Open Source),
  - metadata, annotation schemes, standards,
  - development platforms,
  - evaluation means (adapted to the language specificities, such as for Machine Translation of morphologically-rich languages),
- the effort should be devoted in the long-term, resulting in a necessary strong foundation,
- dialects variants and sociolinguistics should also be taken into account,
- addressing only the short-term development of a specific product or service for that language (as a kind of simple toy), should be avoided, whereas demonstrating applications based on a strong foundation should be favored.
- when a majority language also exists, both should be studied together, and it would save time and efforts to consider a family of languages all together.
- bootstrapping approaches facilitate the coverage of a language.
- cooperation among countries or programs would greatly help by providing the less advanced ones with examples and Best Practices, such as the definition of a commonly agreed basic set of Language Resources which have already been proven necessary to correctly produce the corresponding technologies for a given language, and the identification of gaps and roadmaps should be aimed at.
- the related costs could be shared between the corresponding countries or regions, and international bodies (such as the EC), which could also ensure a proper coordination.
- master keywords should be Interoperability and Sustainability.

2.2 Next meetings

LRL 2011

The 2nd LRL (A JOINT LTC-ELRA-FLaReNet-META_NET EVENT) (2011) was subtitled: "Addressing the Gaps in Language Resources and Technologies".

In the Call for papers the workshop is defined as follows: *"The workshop will draw on the inventories of all language technologies and resources that are presently being carried out, such as the ones conducted by FlaReNet, ELRA or META-NET (e.g. LRE Map, Program Surveys, Language Matrixes, Language Gaps, META-SHARE infrastructure). These are now available and help better understand the current landscape and work out the possible solutions, for each individual language and technology. The idea is to discuss availability, quality, maturity, sustainability, and gaps of the LR and LT for a number of languages and technologies." (www.ltc.amu.edu.pl/2011/)*

This time 15 papers were presented by 30 authors from laboratories in France, India, Ireland, Italy, Japan, Norway, Spain and Switzerland. The contributions addressed the following languages (Basque, Chinese, Irish, Indian languages (12), Khmer, Luxembourgish, Magahi, Punjabi, Quechua, Vietnamese.

LRL 2013

The 3rd LRL WORKSHOP focused on new technologies appearing as a challenge for les resourced languages, as its full name was "A Joint LTC-ELRA-FLaReNet-META-NET Workshop on Less-Resourced Languages): Less Resourced Languages, new technologies, new challenges and opportunities".

[3] Unpublished Report on the Special joint LTC-FLaReNet session « Getting Less-Resourced Languages On-Board!»; LTC'09 Conference Poznan, 6-8 November 2009 by Joseph Mariani, Khalid Choukri and Zygmunt Vetulani.

The theme for this even was defined in the following way: *"Many less resourced languages (LRL) that are thriving to get a place in the digital space and that could profit of the new opportunities offered by the Internet and digital devices will seriously face digital extinction if they are not supported by Language Technologies. Language Technologies (LTs, i.e. spelling and grammar checkers, electronic dictionaries, localized interfaces, voice dictations, audio transcriptions and subtitling, as well as multimedia/multimodal search engines, language translators or information extraction tools) are essential instruments to secure usability of less resourced languages within the digital world, thus ensuring those languages equal opportunities and raising their profile in the eyes of natives but also non-natives from the younger, digitally-oriented generation. However, there are many challenges to be faced to equip less resourced languages with LTs (from basic to advanced): a substantial delay in development of basic technologies, a lack of cooperation among languages communities, a chronic shortage of funding (in particular for minority languages not officially recognized, yet often the most vital ones over the Internet) and the limited economic value placed over LTs for minority languages by the market rules. At this critical time, this workshop seeks to continue the debate as to what new technologies have to offer less resourced languages, and how the research community might seek to overcome the challenges and exploit the opportunities*" *(www.ltc.amu.edu.pl/2013/).*

This time Claudia Soria (CNR-ILC, Italy) joint the group of LRL co-chairs. The workshop attracted 20 authors from 7 countries (France, India, Italy, Norway, Quatar, Spain, USA), who covered by their research over 14 languages (Arabic, Bengali, Chinese, Hindi, Indian languages, Indochinese language, Italian dialects, Italian German, Malay, Malayalam, Khmer, Occitan, Sardinian, Vietnamese) in 9 presentations.

LRL 2015

The 4th LRL Workshop: "Language Technologies in support of Less-Resourced Languages" was co-chaired by Khalid Choukri, Joseph Mariani, Claudia Soria and Zygmunt Vetulani.

Expectations of the LRL co-chairs were articulated as follows:

"This Workshop is targeting all stakeholders somehow involved in Language Technology for less-resourced languages, either as users, developers, researchers, language activists, policy makers. As such, the Workshop broadly addresses current use and usability of Language Technologies for less-resourced languages. This year, we take the opportunity of celebrating the 20th anniversary of the Language and Technology Conference to analyze the influence of Language Technologies on Less-Resourced Languages over two decades. We will particularly welcome contributions addressing the following issues:

1) LRLs in the digital age - how well are regional/minority/less-resourced languages equipped for the digital age? What is the experience of speakers, what are their opportunities to act in the digital sphere by means of these languages? Do speakers of regional/minority/less-resourced languages experience any kind of "unequal digital opportunity"? What is the impact of LRTs on the use and usability of LRL on digital media and devices?

2) LRTs for LRL - development of LRTs for LRLs is often linked to purposes other than availability of applications for retrieving information or for enabling communication (e.g. language learning, identity-building or language reclamation): how often are LRLs targeted by applications for educational, entertainment, or revitalization purposes?

3) LRL: charting the field - what do we know about currently available LRTs for LRL? How to draw a comprehensive and accurate picture? Who are the actors to be involved? What is the experience of researchers and developers?

4) LRL: rethinking the BLaRK - the BLaRK still proves a useful tool for planning and implementing LT for LRL. How can it be remodeled/rethought in the light of current technological development? How can it be channeled into a coherent development roadmap?" (www.ltc.amu.edu.pl/2015/)

The 42 authors of 14 research papers affiliated in 10 countries (Canada, Germany, India, Ireland, Italy, Madagascar, Norway, Poland, Switzerland, UK) contributed to this workshop with their papers concerning over 12 languages (African languages, Ancient Greek, Indian languages, Krio, Malagasy, Sanskrit, Scottish Gaelic, Sambalpuri, Swahili, Swiss German, Vietnamese, Welsh).

LRL 2017

The 5th edition of the Joint LTC-ELRA-FLaReNet-META_NET Workshop on Less-Resourced Languages was announced and prepared by Girish Nath Jha (JNU, New Delhi, India) and Claudia Soria (co-chairs). In order to attract and mobilize the attenders themes and motivation the suggested themes were defined in form of questions:

"LRL: charting the field - what do we know about currently available LTs for LRLs? What is the current status of language technologies and use of LRLs in the digital and social media environments? How to draw a comprehensive and accurate picture and create a road map for future? Who are the actors to be involved? What is the experience of researchers and developers?

LRL: Resource development - how are the LRLs dealing with resource crunch, creation and related issues of standards, IDEs and platforms, funding, usability, sharing etc? What are the perceptions and roles of various stake holders including the governments, industry and language communities? What are the additional challenges posed by multilingual societies? What are the language preservation strategies for LRLs in the digital age?

LRL : technology development - challenges in the development of specific enabling technologies for LRLs at language, speech and multi-modal levels. How are these technologies used in areas such as communication, education, entertainment, health, administration. governance, etc?" (www.ltc.amu.edu.pl/2017/)

The 22 authors of 7 research papers affiliated in 5 countries (Canada, Croatian, France, Japan, Poland, UK) contributed to this workshop with papers concerning languages (Awadhi, Ainu, Braj, Embosi/Bantu, French Vietnamese, Georgian).

LRL 2019

The formula of the workshop was proposed by the new team of co-chairs Dorothee Beermann (Norwegian University of Science and Technology), Laurent Besacier (Grenoble Alpes University, France) and Claudia Soria (CNR-ILC, Italy), following the creation of the joint ELRA-ISCA Special Interest Group on Under-resourced Languages (SIGUL).

This LRL workshop was different from all preceding ones. In particular because, several papers of perfect fit with traditional LRL motivations and objectives, were

accepted to other LTC tracks, mostly because their technical nature. In this number were for example papers from Central Asia, but also Georgia, Iran and Oceania. In order to make this overview complete, we will also take these contribution into account. In total we identified 15 papers whose authors (44) are affiliated in 14 countries (Australia, France, Germany, Georgia, India, Iran, Japan, Kazakhstan, Nigeria, Norway, Thailand, Spain, Uzbekistan, Vanuatu). The following languages, commonly classified less-resourced, were object of studies presented at the conference (LRL and other LTC tracks): Alsatian, Georgian, Efate language, Ibibio, Kazakh, Nefsan language, Oceanian languages, Telugu English, Persian, Thai language, Uzbek, Wolof. (www.ltc.amu.edu.pl)

2.3 Some observations

The first LRL Workshop organized in 2009 was an answer to already identified threats of technological and social exclusion on the global scale due to language barriers and scarcity of communication technologies in the world pretending to become a "global village". The response to the first call for papers in 2009 and further LRL meetings, in particular expansion of the LTC/LRL on countries so far rare at international LT events, confirmed interest in this activity (see table 1 below).

LRL 2009-2019	
Affiliation countries of the authors[4]	29
Number of concerned languages[5]	46
Number of presented papers	68
Number of authors	170

Table 1: LRL in numbers[6].

We observed a number of positive phenomena. Among the LRL papers we identified interesting articles addressing existing needs and reports on research concerning languages rare at the international LT conferences (as some African and Oceanian languages, but also dialects and regional languages in Europe as Occitan and Romani, or, last but not least, historical languages as Ainu or Ancient Greek).

One may expect that any given language, as LT works on this language progress, will not be considered "less-resourced" anymore and will stop being discussed at the LRL workshops. Considering, however, the situation where, according UNESCO, at least 2500 languages in the world is considered endangered, the need of that kind of initiatives will not fizzle out soon.

3. Conclusions

Started 10 years ago, the Less Resourced Languages workshop series has seen both its size and its form undergo various changes depending on the evolution of Natural Language Processing. The huge amount of data, computing power and ubiquitous presence of computers in an ever increasing number of activities requiring innovative applications, combined with the growing economic and cultural pressure of today, make the situation more and more critical for many languages that could disappear in a near future. Their survival lies in part in the existence of events like the LRL workshop series. The initiators and past contributors, we should thank for their support and involvement, hoping that new people will join us in the future in our quest to make Less-Resourced Language become “Better-Resourced Languages".

The last day of the LTC 2019, in our informal meeting already after the conference closure session, we discussed the proposal of establishing an international structure for LT to continue the LTC vocation with particular attention to Less-Resourced and Endangered Languages.

4. Acknowledgements

Thanks to all people who were involved in the workshop series – authors, panelists, PC members, reviewers and debaters – for their contribution to the awareness mission of the LRL events.

5. Bibliographical References

Mariani, J., Paroubek, P., Vetulani, Z. (2009). Report on the Special joint LTC-FLaReNet session « Getting Less-Resourced Languages On-Board!»; LTC’09 Conference Poznan, 6-8 November 2009 (Unpublished)

Mariani, J., Francopoulo, G., Paroubek, P., Vetulani, Z. (2015). Rediscovering 10 to 20 Years of Discoveries in Language & Technology. In Vetulani, Z. and Mariani, J., editors, Proceedings of 7th Language & Technology Conference, Human Language Technologies as a Challenge for Computer Science and Linguistics, Poznań, Poland, ISBN 978-83-932640-8-7, pages 29-47.

Language and Technology Conferences web site (1995-2019). http://www.ltc.amu.edu.pl/.

[4] Australia, Bulgaria, Canada, Croatia, France, Georgia, Germany, Ghana, Greece, Hungary, India, Iran, Irland, Italy, Japan, Kazakhstan, Madagascar, Nigeria, Norway, Poland, Quatar, Serbia, Slovak Republic, Spain, Switzerland, Thailand, UK, Uzbekistan, Vanuatu.

[5] Ainu, Alsatian, Amharic, Ancient Greek, Arabic, Basque, Bengali, Bulgarian, Chinese, Croatian, Early Braj, Embosi(Bantu), Ga language, Georgian, Hindi, Ibibio, Indian Languages, Indonesian, Irish, Kazakh, Khmer, Krio, Kwa languages, Luxemburgish, Magahi, Malagasy, Malay, Malayalam, Nafsan, Occitan, Persian, Pinyin for Taiwanese, Punjabi, Quechua, Romani, Sambalpuri, Sanskrit, Scottish Gaelic, Swahili, Swiss German, Telugu-English, Thai, Uzbek, Vietnamese, Welsh, Wolof.

[6] Figures in Table 1 are underestimated. As the LRL workshop are fully integrated with the rest of LTC, several contributions on low resourced languages were presented in specialized thematic sessions. For LRL 2009-2017 we report here only contributions directly addressed to the LRL and ignore other, submitted to thematic sessions.

Proceedings, pages 127–130
Paris, UNESCO Headquarters, 5-6 December, 2019.

The Linguistic Data Consortium: Developing and Distributing Language Resources4All

Denise DiPersio, Christopher Cieri
Linguistic Data Consortium, University of Pennsylvania
3600 Market Street, Suite 810, Philadelphia, PA 19104 USA
dipersio@ldc.upenn.edu, ccieri@ldc.upenn.edu

Abstract

The Linguistic Data Consortium (LDC) is an open collective of academic, government and industry organizations whose mission is to support language-related research, education and technology development by creating and sharing resources, such as data, tools and standards. Its online catalog is a rich, curated repository of speech, text, video and lexical data sets. LDC develops and publishes resources in a growing number of underserved languages. This paper examines relevant LDC corpora and the "language pack" data set model as successes in resource creation, along with the Consortium's involvement in efforts that advance access to data for all language communities.

Keywords: language resources, digital repositories, language communities

Résumé

Ilé-iṣẹ́ Àgbájọ Fún Àkójọpọ̀-ọ̀rọ̀-èdè Ajẹmọ́-ìmọ̀-ẹ̀dá-èdè (Linguistic Data Consortium (LDC)) jẹ́ ilé-iṣẹ́ àgbájọ àwọn onímọ̀-ìjìnlẹ̀, àwọn ìjọba àti àwọn ilé-iṣẹ́ ìmọ̀ ẹ̀rọ, pẹ̀lú èròǹgbà láti ṣe àtìlẹ́yìn fún ìdàgbàsókè ìjìnlẹ̀ ìwádìí ajẹmọ́-èdè, ètò ẹ̀kọ́, àti ìmọ̀ ẹ̀rọ, nípa pípèsè àti pípín àkójọ-èdè, gẹ́gẹ́ bí àkójọ-ọ̀rọ̀ nínú èdè, àwọn oríṣiríṣi ohun-èèlò, àti àwọn ìlànà ìgbéléwọ̀n. Ààtò àgbéjáde rẹ̀ nínú ẹ̀rọ ayélujára kún fún àkópamọ́ ọ̀rọ̀-ẹnu geere, ọ̀rọ̀-àkọsílẹ̀ oníhun, àti oríṣiríṣi àká-ọ̀rọ̀. Ilé-iṣẹ́ yìí (LDC) ń ṣe iṣẹ́ ìdàgbàsókè pẹ̀lú àwọn àtẹ̀jáde, lórí ìpèsé ohun àmúlò fún ọ̀pọ̀lọ́pọ̀ àwọn èdè kéékèèkéé tí wọ́n ń dàgbàá bọ̀. Ìwé àpilẹ̀kọ yìí ń ṣe àlàyé lórí àwọn àkójọpọ̀ ọ̀rọ̀-èdè ti LDC àti "àṣàjọ ìwé" fún ọ̀rọ̀-èdè, gẹ́gẹ́ bí àwòkọ́ṣe lórí àwọn àṣeyọrí nínú ìṣẹ̀dá ohun àmúlò, pẹ̀lú ojúṣe LDC nínú akitiyan láti mú kí àkójọpọ̀ ọ̀rọ̀-èdè fún gbogbo àwùjọ ajọlèdè wà ní àrọ́wọ́tó. ti awọn ede ti ko ni idaniloju. Iwe yii ṣe ayẹwo corpora LDC ti o yẹ ati awoṣe "idii ede" awoṣe apẹẹrẹ bi awọn aṣeyọri ni ẹda awọn orisun, pẹlu ikopapọ Consortium ninu awọn ipa ti o ni ilosiwaju si data fun gbogbo agbegbe agbegbe.

1. Introduction

This paper introduces the Linguistic Data Consortium (LDC), an open consortium of universities, libraries, corporations and government research laboratories hosted at the University of Pennsylvania USA and describes how it fulfills its mission to develop and broadly share language resources. LDC's online catalog is a rich, curated repository of multilingual speech, text, video and lexical data sets that includes publications of interest to underserved language communities. The Consortium also works with like-mended global sister organizations and networks to advance language-related research, education and technology development in the world's languages.

2. LDC: Founding, Mission and Operation

LDC was founded in 1992 to address the critical data shortage then facing language technology research and development on the principle that broad access to data drives innovation. Its mission is to support language-related education, research and technology development by creating and sharing linguistic resources, such as data, tools and standards. From its primary role as a repository and distribution point for language resources, the Consortium has grown into an organization that creates and distributes a wide array of language resources to the global community and supports sponsored research programs and language-based technology evaluations by providing resources and contributing organizational expertise.

The Consortium is a mutual aid society. Researchers contribute data sets to the LDC Catalog and as a result, their work gains visibility and community recognition and inspires other research. Members and data licensees contribute fees and in return receive ongoing rights to a variety of resources; those fees are typically a fraction of the cost of data development. Sponsors contribute funding that results in resource creation, infrastructure, innovation, cost sharing and resource dissemination to the community.

3. The LDC Catalog

3.1 Sharing Data in the World's Languages

The LDC Catalog is a growing digital archive of over 800 holdings that for more than two decades has served as one of the world's major language resource repositories. As the first and most active language resource data center, LDC established or adopted many of the publication, archiving and curation practices that related research communities follow today. Originally seeded by data contributions of significant corpora, the catalog continues to be augmented by data sets developed by LDC and by donations from researchers worldwide. As of this writing, LDC has distributed close to 200,000 copies of its resources in over 90 languages to roughly 6000 distinct organizations in more than 100 countries. Over 10,000 unique papers citing LDC data have been identified, attesting to the repository's overall research impact.

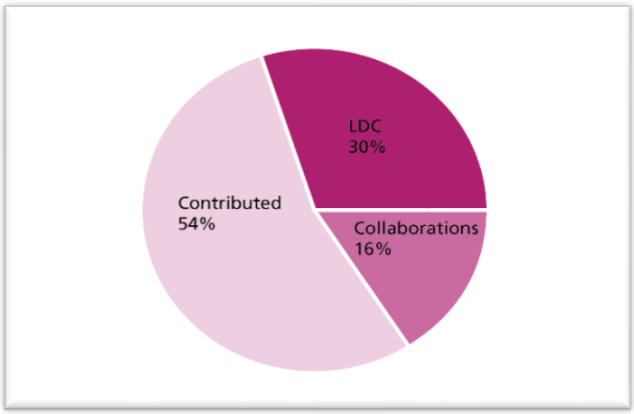


Figure 1: The LDC Catalog is a community resource.

3.2 Curating Language Resources

The catalog has also been recognized as a trustworthy data repository under the CoreTrustSeal certification established by the ISCU World Data System and the Data Seal of Approval. [1] This means that the Catalog meets high standards for data access, rights management, curation, data integrity and authenticity, archival storage and security. The Catalog also consistently receives the highest (five-star) rating for compliance with the Open Language Archives Community (OLAC) metadata standard, an extension of the Dublin Core standard designed for language resources.[2]

LDC's curation workflow includes data review upon submission, a battery of quality checks, metadata creation and documentation development. The data is then prepared for delivery, usually via web download or on media for larger corpora. All data distributed through the catalog is archived in a logical data tree subject to a specialized back-up system from which it can be migrated to new formats, platforms and storage media as required by best practices in the digital preservation community.

LDC's licenses are compatible with the community's customary uses as well as with intellectual property, human subjects and privacy concerns. These include tribal rights in community languages, recently reaffirmed in revised US human subjects regulations.

4. Language Resources Overview

LDC develops and publishes resources in a growing number of languages referred to under several terms: indigenous languages, minority languages, endangered languages and low resource languages. Whatever the name, such languages pose challenges to researchers. Human language technology development relies on digital resources, such as lexicons, grammars, monolingual and parallel corpora, morphological analyzers, taggers and segmenters. For some languages, the source data is scarce; for others the structure of the language itself affects the development of technology-related resources. Below is an overview of some LDC data sets and research noting solutions to language-specific issues.

4.1 West African Languages

Among the research challenges presented by West African languages are complex phonology and morphology (Bantu), verb serialization (Kwa), complex pronoun systems (Yoruba) and the absence of established writing systems (many). LDC data sets in the Manding languages, Yoruba and Dschang and Ngomba (Bantu) illustrate creative and flexible solutions to language challenges.

Grassfields Bantu Fieldwork: Dschang and Ngomba Tone Paradigms. Tonological and phonetic description of tone paradigms.
Global Yoruba Lexical Database. Diaspora dialects included to capture the language's global impact: Nigeria and Benin to the Caribbean and islands along the southeastern United States coast.
Manding lexicons (Bamanankan, Maninkakan, Mawukakan). Bidirectional English and French glosses to accommodate speakers in a francophone context.

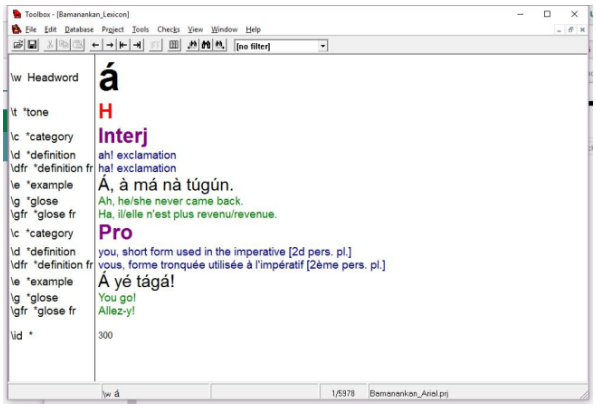


Figure 2: Entry from Bamanankan Lexicon (LDC2016L01) in Toolkit interface

4.2 Fieldwork

Some recent approaches in fieldwork documenting endangered languages incorporate simple technologies like handheld recorders and smartphones to allow large numbers of community members to capture speech for re-speaking, transcription and translation. LDC contributed to two such studies in **Papua New Guinea** and **Brazil** funded by the National Science Foundation (BCS-0951651, IIS-0964556).

Malto is a Dravidian language spoken in northeastern India and Bangladesh by people called the Parahiyas in villages or hamlets located on hilly tracts and in the lowlands. The **Malto Speech and Transcripts** corpus contains audio data from speakers who share their life stories, local rituals from

[1] https://www.coretrustseal.org/.

[2] http://www.language-archives.org/.

festivals to funerals, and the oral histories and rich folklore of their community.

4.3 Language Packs

LDC has developed « language packs » for low resource languages in two US government-funded projects, REFLEX and LORELEI. The idea behind the language pack is to construct a core set of language resources and tools that can be deployed for multiple purposes, among them, language documentation and preservation, basic technology development and situational awareness, e.g., natural and humanitarian disasters. (Simpson, et al., 2008 ; Strassel and Tracey, 2016).

Figure 3: Collaborative transcription in Papua New Guinea (Courtesy: Steven Bird)

Language packs consist of monolingual text, parallel text, several types of annotation, tools for text processing, segmentation and entity tagging, as well as lexicons and grammatical sketches. Languages covered include **Akan (Twi), Amazigh, Amharic, Ilocano, Kinyarwanda, Odia, Oromo, Sinhala, Tigrinya, Uighur, Uzbek, Wolof and Zulu.** It is expected that packs for roughly 20 languages will be released into the LDC Catalog beginning in 2020.

5. Research Collaborations in Indigenous Languages

LDC is an active participant in related consortia and groups whose aim is to advance the ways in which resources are developed and distributed. These include initiatives for standardizing specifications and best practices and for developing new architecture to support language resource delivery. Collaborations involving indigenous languages are among these.

5.1 Community Projects

In the National Science Foundation's **AARDVARC project** (Automatically Annotated Repository of Digital Audio and Video Resources Community), [3] LDC engaged with an interdisciplinary community of linguists, anthropologists and computer scientists to discuss and develop standards around formats, access and use of resources in endangered and low resource languages.

Similarly, in **E-MELD** (Electronic Metastructure for Endangered Languages Data), [4] LDC participated in the effort to develop consensus on documenting endangered languages and fostering collaboration among digital archives.

5.2 Languages of the Americas

LDC promotes resource development in the Americas in a variety of ways. These include advice and technical assistance for specific collections, among them, Nahuatl and Mixtec. Recently, the Consortium convened two workshops in 2018 exploring hemispheric collaboration and language resource development.

The **Planning Workshop on Data Archives and Languages of the Americas**[5] was held in Philadelphia with support from the University of Pennsylvania. Experts managing linguistic data archives and resource centers met to discuss challenges, needs and opportunities for promoting and extending collaboration in the Americas.

The **International Workshop on Data Intensive Research on Languages in the Americas,**[6] also supported by the Penn Global initiative at the University of Pennsylvania, took place in Mexico City. Linguists and scientists from Mexico, Brazil, Chile, Argentina and the United States presented their work on Chuj, Yucateco, Huasteco, Nahuatl, Wixarika, Southern Cone languages, Mexican/American Spanish and Brazilian Portuguese.

These collaborations have provided the beginnings of a strong regional community and the basis for future work.

Figure 4: LDC's global network of contributors and collaborators

6. Innovation in Language Resource Development

Despite the large volumes of linguistic data created by current methods, supply continues to lag far behind demand. This is due in part to the application of a finite resource to a problem that is effectively infinite or at least several orders of magnitude larger.

LDC has recently begun to address this problem by identifying renewable sources of the time and intellectual investment required to document the world's languages, especially for the purposes of human language technology development. The experience of social media platforms, grass roots efforts such as Librivox, which creates

[3] https://www.nsf.gov/awardsearch/showAward?AWD_ID=1519887.

[4] http://emeld.org/.

[5] https://www.ldc.upenn.edu/communications/workshops/penn-urf-sas-workshop.

[6] https://www.ldc.upenn.edu/communications/workshops/penn-gef-americas-workshop.

audiobooks from out-of-copyright texts, and especially citizen science platforms, demonstrates that the human drive for challenge, advancement, entertainment and the opportunity to contribute to one's own betterment and that of one's local community and the broader society are effectively boundless. For example, nearly two million contributors to the Zooniverse citizen science portal have submitted more than 250 million judgments that are used by researchers in astronomy, biology and other fields.

LDC's *LanguageARC* presents language resource projects to potential Citizen Linguists. Each includes multiple tasks that require a simple judgment repeated over multiple items. For example, one project might seek to document the state of a number of indigenous languages of South Africa through surveys that document the point in children's development at which they acquire the words for culturally significant objects. Another might elicit local terms via picture or silent video description. Still others might elicit re-speaking or translations as a way to reveal the grammatical features of a language. LanguageARC supports any task in which contributors are shown a text or images or are played audio or video clips and asked to respond to instructions that are either specific to the task or that vary with each item by speaking, entering a text response or selecting one or more items from a multiple choice list.

To attract and support a community of contributors each LanguageARC project has a title, call to action, image, pitch, picture, partner badges, description of the research team and discussion forums to support community building. Tasks similarly have a title, calls to action and images but also include tutorials, reference guides and their own discussion forum.

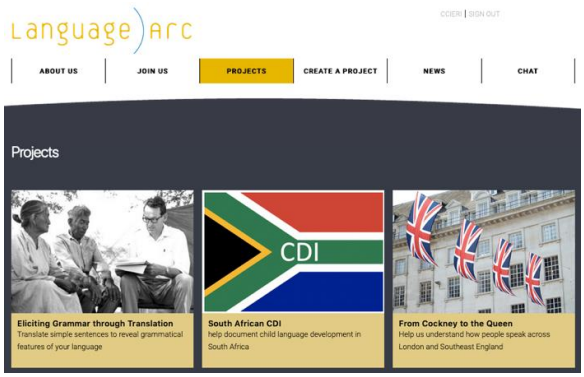


Figure 5: The LanguageARC Citizen Linguist portal

7. Conclusion

Access is a crucial theme of the 2019 International Year of Indigenous Languages – access to education, information and knowledge for indigenous peoples in their home languages. Means to that end include the availability of data collections that in turn can be used to develop language technologies for indigenous language communities. LDC's founding principle that broad access to data drives knowledge and research resonates with that theme. The Consortium is committed to developing and sharing resources in all languages for all language communities in ways that ensure meaningful access, advance language vitality and promote preservation.

8. Bibliographical References

Bamanankan Lexicon. (2016). Distributed via LDC, LDC2016L01, ISLRN 830-816-122-814-4.

CoreTrustSeal. https://www.coretrustseal.org/. Accessed 25 November 2019.

EMELD. http://emeld.org/, Accessed 25 November 2019.

Federal Policy for the Protection of Human Subjects. 82 Fed. Reg. 7149 (Jan. 19, 2017).

Global Yoruba Lexical Database v 1.0 (2008). Distributed via LDC, LDC2008L03, ISLRN 973-344-578-516-8.

Grassfields Bantu Fieldwork: Dschang Lexicon. (2003). Distributed via LDC, LDC2003L01, ISLRN 880-081-036-797-6.

Grassfields Bantu Fieldwork: Dschang Tone Paradigms. (2003). Distributed via LDC, LDC2003S02, ISLRN 973-117-906-652-9.

Grassfields Bantu Fieldwork: Ngomba Tone Paradigms. (2001). Distributed via LDC, LDC2001S16, ISLRN 147-689-240-962-1.

International Workshop on Data Intensive Research on Languages of the Americas. https://www.ldc.upenn.edu/communications/workshops/penn-gef-americas-workshop. Accessed 25 November 2019.

Malto Speech and Transcripts. (2012). Distributed via LDC, LDC2012S04, ISLRN 841-757-472-203-8.

Maninkakan Lexicon. (2013). Distributed via LDC, LDC2013L01, ISLRN 573-342-913-646-6.

Mawukakan Lexicon. (2005). Distributed via LDC, LDC2005L01, ISLRN 592-356-503-307-6.

National Science Foundation. Collaborative Research: Automatically Annotated Repository of Digital Video and Audio Resources Community (AARDVARC). https://www.nsf.gov/awardsearch/showAward?AWD_ID=1519887. Accessed 25 November 2019.

OLAC: Open Language Archives Community. http://www.language-archives.org/. Accessed 25 November 2019.

Planning Workshop on Data Archives and Languages of the Americas. https://www.ldc.upenn.edu/communications/workshops/penn-urf-sas-workshop. Accessed 25 November 2019.

Simpson, H., et al. (2008). Human Language Technology Resources for Less Commonly Taught Languages: Lessons Learned Toward Creation of Basic Language Resources. In Nicoletta Calzolari (Conference Chair), et al., editors, Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), W10, Collaboration: interoperability between people in the creation of language resources for less-resourced languages, pages 7-11, Marrakesh, Morocco, May. European Language Resource Association (ELRA).

Strassel, S. and Tracey, J. (2016). LORELEI Language Packs: Data, Tools, and Resources for Technology Development in Low Resource Languages. In Nicoletta Calzolari (Conference Chair), et al., editors, Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 3273-3280, Reykjavik, Iceland, May. European Language Resource Association (ELRA).

Proceedings, pages 131–136
Paris, UNESCO Headquarters, 5-6 December, 2019.

European Language Grid: Language Technologies for Europe

Georg Rehm
DFKI GmbH, Germany
georg.rehm@dfki.de

Abstract

With 24 official EU and many additional languages, multilingualism in Europe and an inclusive Digital Single Market can only be enabled through Language Technologies (LTs). European LT business is dominated by hundreds of SMEs and a few large players. Many are world-class, with technologies that outperform the global players. However, European LT business is also fragmented – by nation states, languages, verticals and sectors, significantly holding back its impact. The European Language Grid (ELG) project addresses this fragmentation by establishing the ELG as the primary platform for LT in Europe. The ELG is a scalable cloud platform, providing, in an easy-to-integrate way, access to hundreds of commercial and non-commercial LTs for all European languages. The ELG will enable the commercial and non-commercial European LT community to deposit and upload their technologies and data sets, to deploy them through the ELG, and to connect with other resources.

Keywords: LR Infrastructures and Architectures, LR National/International Projects, Tools, Systems, Applications, Web Services

1. Introduction

With 24 official EU languages and many additional ones, multilingualism, cross-lingual and cross-cultural communication in Europe as well as an inclusive EU Digital Single Market can only be enabled and firmly established through Language Technologies (LTs) (Rehm, 2016). The European LT industry is dominated by hundreds of SMEs and a few large players. Many are world-class, with technologies that outperform the global players. However, European LT business is also fragmented – by nation states, languages, domains and sectors (Vasiljevs et al., 2019), significantly holding back its impact. In addition, many European languages are severely under-resourced and, thus, in danger of digital language exinction (Rehm and Uszkoreit, 2012; Kornai, 2013; Rehm et al., 2014; Rehm et al., 2016a), which is why there is an enormous need for a European LT platform as a unifying umbrella (Rehm and Uszkoreit, 2013; Rehm et al., 2016b; STOA, 2017; Rehm, 2017; Rehm and Hegele, 2018; European Parliament, 2018).[1] The project European Language Grid (ELG; 2019-2021) addresses this fragmentation by establishing the ELG as the primary platform and marketplace for the European LT community, both industry and research. The ELG is a scalable cloud platform, providing access to hundreds of commercial and non-commercial LTs for all European languages, including running tools and services as well as data sets and resources. The ELG will enable the commercial and non-commercial European LT community to upload their technologies and data sets into the ELG in an easy and efficient way, to deploy them, and to connect with other resources.

2. Approach and Methodology

The European LT community has been demanding a dedicated LT platform for years. The ELG project's ambition is to establish the European Language Grid as the primary platform for industry-relevant LT in Europe, bringing together and uniting a network of European experts and concentrating on *commercial* and *non-commercial LTs*, both *functional* and *non-functional* (corpora, lexicons, data sets etc.). A related goal is to establish the ELG as the primary market place for the fragmented European LT landscape to connect demand and supply, strengthening Europe's position in this field. The ELG is meant to enable the whole European LT community to upload their services and data sets, to deploy them and to connect with, and make use of those resources made available by others. *The ELG is meant to be a shared platform for the whole European LT community*, enabling LT provider companies to grow and benefit from scaling up and also companies who want to integrate LT into their products or services.

[1] This article is a shortened version of (Rehm et al., 2020a).

3. The European Language Grid

In the following, we describe the architecture of the ELG platform (Section 3.1.), including the repository catalogue (Section 3.2.) as well as the user interface (Section 3.3.). Section 3.4. provides more details on the functional services available in the ELG. The language resources are discussed in Section 3.5., followed by a description of the ELG community (Section 3.6.). Section 3.7. provides an overview of the open calls for pilot projects.

3.1. Technical Architecture of the Infrastructure

ELG is a scalable platform with an interactive web user interface and backend components and APIs. It offers access to various language resources as well as functional LT services, i. e., LT tools that have been containerised and wrapped with the ELG LT Service API. ELG's integrated functional services can be used through APIs or through the web interface. The architecture is separated into three layers (Figure 1), i. e., the *base infrastructure* (Kintzel et al., 2019; Moritz et al., 2019), the *platform backend* (Piperidis et al., 2019; Labropoulou et al., 2019) and the *platform frontend* (Melnika et al., 2019a; Melnika et al., 2019b).

The *base infrastructure* is operated on a Kubernetes cluster in the data centre of a European provider located in Berlin, Germany. All infrastructural components run as Docker containers in this cluster.

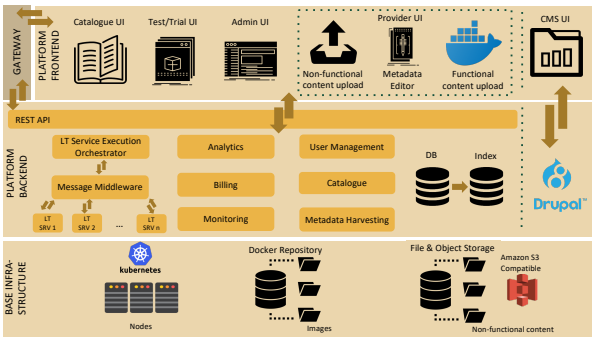


Figure 1: Technical architecture of the ELG

The *platform backend* contains the ELG catalogue, i. e., the metadata records of functional services, non-functional resources but also the entries of organisations (e. g., companies, universities) and other stakeholders, as well as service types, languages and other types of information. Stakeholders will be able to register themselves in this catalogue for increased reach and visibility. Users will be able to filter and search for organisations, services, data sets and more, by languages, service types, domains, and countries. The platform backend layer also includes the LT Service Execution Server/orchestrator that offers a common REST API for calling integrated functional services.

The *platform frontend* layer consists of UIs for the different types of ELG users, e. g., LT providers, potential buyers and ELG system administrators (Section 3.3.). These include catalogue UIs, test/trial UIs for functional services, provider UIs for uploading/registering functional services etc.

One of the key concepts of the architecture is the use of containers to encapsulate all components, settings and libraries of an individual LT service in one self-contained unit. Docker is currently the most widely used technology for containerisation. For individual LT services, Docker images can be built locally by their respective providers and ingested into the ELG, where they can be started, terminated and scaled out on demand.

3.2. Catalogue Structure – Metadata Schema

The ELG catalogue contains all entities of interest to users (Section 3.6.), appropriately indexed and described so that they can easily find and select the resources that meet their requirements and deploy them, as well as visualise the LT domain activities, stakeholders and resources with specific criteria (e. g., service type, language, etc.). All entities are described in compliance with the ELG-SHARE metadata schema (Labropoulou et al., 2019; Labropoulou et al., 2020). The schema builds upon, consolidates and updates previous activities, especially the META-SHARE schema and its profiles (Gavrilidou et al., 2012; Piperidis et al., 2018; Labropoulou et al., 2018), taking into account the ELG user requirements (Melnika et al., 2019a), recent developments in the (meta)data domains (e. g., FAIR, data and software citation recommendations, Open Science movement, etc.), and the need for establishing a common pool of resources through exchange mechanisms with collaborating projects and initiatives (Section 3.6.).

3.3. User Interface

To identify the user scenarios and requirements, we defined the main groups of ELG users: (1) *Content providers* – companies, research organisations or public institutions with tools, services, or data that can be provided through the ELG; (2) *Developers and integrators* – companies and research institutions interested in using LT services, tools, or data in their applications; (3) *Information seekers* – users interested in information about LT, data or events; (4) *Information providers* – organisations or individuals who wish to provide information about events, trainings etc.; (5) *Casual visitors*; (6) *ELG platform administrators*.

Angular and Typescript are used for developing the ELG front-end. The Angular Material components are implemented as an adjustable theme that can be tuned to the designer's specifications. For example, a theme has primary and secondary colours that, once set, will be used throughout all interface elements. The website design is based on the Single Page Application (SPA) principle. To enable the flexible management of content and information within the ELG website, we integrated Drupal. As the ELG front-end is a monolithic SPA, the CMS does not have a dedicated public front-end. Instead, Drupal serves different menus and page contents using REST services and JSON-HAL.

3.4. Functional Services

The European LT market is very broad and varied, with many different providers of many different classes of services and tools, exposed as many different APIs and data formats. One of ELG's primary goals is to attempt to bring more order to this varied landscape by identifying classes of related services and providing generic APIs for each class. From the outset the project has identified a number of broad classes that share much in common:

Machine Translation (MT): services that take text in one language and translate it into text in another language, possibly with additional metadata associated with each segment (sentence, phrase, etc.). This class can include (seemingly unrelated) services such as summarisation, where the summary can be viewed as a "translation" of the original text.

Information Extraction (IE): services that take text and annotate it with metadata on specific segments. This class can cover a wide variety of services from basic NER through to complex sentiment analysis and domain-specific tools.

Automatic Speech Recognition (ASR): services that take audio as input and produce text (e. g., a transcription) as output, possibly with metadata associated with each segment.

Other clusters are emerging as the project considers more services for integration, for example text-to-speech, text classification, alignment, and translation quality estimation. An aspiration for the platform is to provide services of all classes for all official EU languages and for other EU and non-EU languages that are of strategic interest within the EU. The current prototype has so far integrated seven IE and text analysis tools plus a multilingual dependency parser supporting 60 languages, five ASR services (one supporting three languages and another supporting two), 14 MT services (six languages into English, English into eight other languages) and text-to-speech in four languages.

3.5. Data Sets and Language Resources

The ELG consortium has defined an LR identification and sharing strategy. It starts by liaising with and capitalizing on existing activities to ingest LRs into the ELG, which often requires some sort of negotiation with the owners/providers to obtain the rights to do so. We currently focus on providers who are part of the consortium (ELDA/ELRA) and recent well-known activities such as ELRC-SHARE (Lösch et al., 2018; Piperidis et al., 2018) and META-SHARE (Piperidis, 2012; Piperidis et al., 2014). ELG is working both on data integration procedures, where metadata compliance is key for the exchange of data and metadata descriptions (Section 3.2.), and on the implementation of market place-related features, such as upload/download, licensing, billing, payment, etc.

LR modalities covered are text (corpora, lexicons, etc.), speech/audio, video/audiovisual, images/OCR, sign language, and others. We currently work on a classification matrix that includes LR types, modalities and languages with the goal of analysing the status of existing LRs and LTs. About 220 additional repositories have been located so far, which will increase the current numbers as the exploration and ingestion of LRs is progressing. ELG will approach resource users and suppliers to offer them an additional market channel and will look into both research organisations and companies that build or use commercial or non-commercial LRs. As a first step, over 650 LRs from ELRA, ELRC-SHARE and META-SHARE are being prepared for integration into the ELG, to be completed for the first release of the ELG platform (April 2020). The following resources will be comprised.

3.6. Stakeholders and Community

ELG aims to respond to the challenge of Europe's fragmented European LT landscape (Vasiljevs et al., 2019), both with regard to industry and research. We address this issue by bringing together all stakeholders under a common umbrella platform, which is why outreach, communication and further community building play a crucial role in ELG. Our main target users are described in Section 3.6.1. In addition, we have been setting up two community instruments, the National Competence Centres (Section 3.6.2.) and the European LT Council (Section 3.6.3.).

3.6.1. Key Stakeholders

ELG caters, first, for *commercial LT providers* who want to showcase their products, services and their own organisation. We want to provide *the* marketplace for European LT, which requires a broad geographical, technological and sectorial representation of, ideally, all relevant European provider companies. Organisations will be able to claim (or delete) their record through the ELG user interface so that they can take over maintenance and populate their ELG page. *Research centres and universities* are also LT providers but their interest is not a monetary but a research-driven one. This group provides data sets or smaller tools including rudimentary, experimental services that have evolved from research projects, rather than fully-fledged services ready for production and monetisation. *LT users* are the most diverse target group. It includes organisations who want to make use of LT, students doing research for a paper or job seekers. Members of this group can be on the lookout for information, try to find free services or be potential buyers. ELG is collaborating or in the process of setting up collaborations with several relevant projects and initiatives that have similar goals, such as AI4EU, ELRC, BDVA, CLAIRE, CLARIN, HumanE-AI and META-NET (Rehm et al., 2020b). The *participants in the pilot projects*, funded through the ELG open calls, are also key stakeholders. ELG will test the platform and demonstrate its usefulness with the help of 15-20 pilot projects that receive financial support through the project budget. The results of these pilots will be fed back into the ELG platform.

3.6.2. National Competence Centres

ELG set up 32 National Competence Centres (NCCs) to extend the reach of the platform and initiative. They were selected based on their involvement in relevant community initiatives. The fact that all NCCs have good connections to major local industry sectors while being part of academic organisations, guarantees independence from economic interests while ensuring sufficient outreach into commercial fields to serve the purpose of ELG. The NCCs function as bridges between the national and regional markets and the ELG, both as a platform and project. They provide information about stakeholders, services, data sets, resources and technologies from the given region. They know the language(s) and the political as well as economic situation in their countries and are represented in regional networks.

3.6.3. European LT Council

ELG is also initiating a second new body, the European LT Council (LTC), as a pan-European group in which strategic LT-related matters can be discussed and coordinated. While the main purpose of the NCCs is to support the mission of the ELG project, the main goal of the LTC is to support and represent the European LT community. The LTC is meant to be a forum that enables easy and efficient communication and coordination at the European level, specifically with regard to ongoing and emerging international and also national activities relating to LT research, development and innovation. The LTC fosters the coordination and strategic as well as political discussion, representing all relevant stakeholder groups. It will prepare strategic recommendations, especially geared towards national and European administrations and funding agencies.

3.7. Open Calls for Pilot Projects

ELG will provide close to 30% of its overall budget for a set of 15-20 small scale demonstrator pilot projects in the form of grants awarded after a call for proposals. The pilot projects will broaden ELG's portfolio by developing missing services or solutions that support underrepresented languages. At the same time, they will demonstrate the ELG's usefulness as a technology platform, especially with regard to sectors of high commercial or societal impact. The results of the projects will be made available through the ELG. LT tools or services will be integrated into the ELG itself and made generally available under defined licensing conditions. The main objective of the open calls is to support SMEs that have long-term potential to either (a) contribute

services, tools or data sets to the ELG to increase its coverage or (b) develop applications using LTs available in the ELG. Financial support will be awarded to selected applicants following an open, transparent and expert-evaluation based selection process. Each proposal will be evaluated by three independent experts for the following criteria: (a) objective fit; (b) technical approach; (c) business, integration and dissemination plan; (d) budget adequacy; and (e) team. The first call will be published in March 2020, and the second in September 2020.

4. Sustainability through a Legal Entity

Achieving the intended scale of the ELG requires a high availability and performance of the overall system, service level agreements (SLAs) for the services as well as billing and support facilities. These create various costs, that can only be covered adequately through a sustainable, long-term operational model. Costs include cloud hosting and bandwidth, personnel costs for operations, development, accounting, marketing, support and management, legal consulting (SLAs, GDPR, contracts etc.), office space, computers, electricity etc. ELG is meant to be a sustainable activity. To achieve this goal, we need to identify ways to cover the incurred costs on a long-term basis. We will establish a legal entity by approx. Q1/2021. Among the options are a for profit or non-profit company, a professional stakeholder association and a foundation.

There are various potential ingredients of a future ELG business and operations plan. These include online ads (for companies, services, conferences etc.), sponsored content (e. g., first search result, clearly marked as "sponsored"), i. e., sponsored services, data sets, or companies, among others. The ELG legal entity can also offer training events, tutorials or webinars for a fee for commercial players, while keeping them free for academia. ELG conferences may include registration fees for delegates from industry, also offering sponsorship packages for companies. Consulting services around ELG and language-centric AI can be offered. If we decide to establish a professional business association, membership fees could be part of the business plan. Project grants can be used to sustain part of the operation. Additionally, the hosting of commercial LT services, models or data sets can be associated with a certain fee, while the results of publicly funded research can be made available for free, but the hosting costs would need to be covered nonetheless. In that regard, ELG could function as the secondary or maybe even as the primary dissemination channel for research projects or for companies that develop LT. Part of the ELG business model could also include the brokering of commercial LT services for a fee, with a split between the service owner and ELG as the broker. ELG could also function as a paid hoster for whole service or data repositories.

5. Related Work

Research Projects, Platforms, Initiatives All in all, we have collected more than 30 projects, platforms and initiatives that are, in one way or another, relevant for ELG (Rehm et al., 2020a). They share at least one of the following goals with ELG: 1) they provide a collection of LT/NLP tools or data sets; 2) they provide a unified platform, which, underneath, harvests metadata records of data sets or services or tools from distributed sources; 3) they provide a sharing platform for the exchange of tools or data sets among stakeholders.

Global Technology Enterprises Many of the global technology enterprises offer a wide range of different processing services, beyond language, including cloud and compute resources, storage, different types of databases, data analytics, and also more engineering-related services such as encryption, development and deployment. Among these are offerings by Amazon, especially AWS and Comprehend, Microsoft Azure Cognitive Services (Del Sole, 2018), the Google Cloud Platform and the IBM Cloud (Kochut et al., 2011). Google has recently (Sept. 2018) released a dedicated search platform for datasets.

6. Conclusions and Next Steps

It has repeatedly been argued that Europe should by no means outsource its multilingual communication and language challenge to providers from other continents since the European demands are so unique and complex (Rehm and Uszkoreit, 2013; Rehm, 2017; Rehm et al., 2020b). Instead, Europe should make use of its own excellent LT community. One of the obstacles to be overcome along the way is the creation of a shared platform for the whole community. The ELG will foster language technologies *for Europe* built *in Europe*, tailored to our languages and cultures and to our societal and economical demands, benefitting the European citizen, society, innovation and industry. There is currently no other scalable cloud platform that can play the role as a joint marketplace and broker for such a broad variety of services and data sets as we have foreseen for the ELG.

At the time of writing, the three-year ELG project is at the end of its first year, which has already seen the first public demo of a fully functional miminum viable product of the ELG platform at META-FORUM 2019. Work in all three ELG areas is progressing at a fast pace. The next major milestones will include launching the first open call in March 2020 and, at the same time, making available the first version of the ELG platform to interested parties. This version will also include the first batches of functional services and data sets. The second open call will be published in September 2020, coinciding with the second release of the platform, services and data sets. The third and final release of the platform (including additional services and data sets) is foreseen for the last quarter of 2021. In 2020 and 2021 we will organise two more annual ELG conferences that will also include NCC and LTC meetings. At the end of 2021, a new legal entity will take over the further development and maintenance of the ELG platform. With regard to upcoming funding programmes on the European level, we foresee ELG to play a number of roles, especially as the main data and service provision and dissemination platform for the European LT and language-centric AI community (Rehm et al., 2020b) in Horizon Europe and Digital Europe Programme but also in national funding initiatives.

Funding notice

The work presented in this paper has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 825627 (European Language Grid) and from the German Federal Ministry of Education and Research (BMBF) through the project QURATOR (Wachstumskern no. 03WKDA1A).

7. Bibliographical References

Del Sole, A. (2018). Introducing Microsoft Cognitive Services. In *Microsoft Computer Vision APIs Distilled*, pages 1–4. Springer.

European Parliament. (2018). Report on language equality in the digital age. http://www.europarl.europa.eu/doceo/document/A-8-2018-0228_EN.html, September. (2018/2028(INI)). Committee on Culture and Education (CULT), Committee on Industry, Research and Energy (ITRE); Rapporteur: Jill Evans.

Gavrilidou, M., Labropoulou, P., Desipri, E., Piperidis, S., Papageorgiou, H., Monachini, M., Frontini, F., Declerq, T., Francopoulo, G., Arranz, V., and Mapelli, V. (2012). The META-SHARE Metadata Schema for the Description of Language Resources. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*. European Language Resources Association (ELRA).

Hummel, J., Wetzel, M., Kranias, L., and Magnusdottir, G. (2016). LTI Cloud Business Model Report, April. MLi Deliverable D4.3. MLi: Towards a MultiLingual Data & Services Infrastructure.

Kintzel, F., Moritz, M., and Rehm, G. (2019). Requirements and Architectural Specification of the Base Infrastructure, April. ELG Deliverable D1.1. ELG: European Language Grid.

Kochut, A., Deng, Y., Head, M. R., Munson, J., Sailer, A., Shaikh, H., Tang, C., Amies, A., Beaton, M., Geiss, D., et al. (2011). Evolution of the IBM Cloud: Enabling an enterprise cloud services ecosystem. *IBM Journal of Research and Development*, 55(6):7–1.

Kornai, A. (2013). Digital Language Death. *PLoS ONE*, 8(10). https://doi.org/10.1371/journal.pone.0077056.

Labropoulou, P., Galanis, D., Lempesis, A., Greenwood, M., Knoth, P., Eckart de Castilho, R., Sachtouris, S., Georgantopoulos, B., Martziou, S., Anastasiou, L., Gkirtzou, K., Manola, N., and Piperidis, S. (2018). OpenMinTeD: A Platform Facilitating Text Mining of Scholarly Content. In *WOSP 2018 Workshop Proceedings, Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 7–12, Miyazaki, Japan. European Language Resources Association (ELRA).

Labropoulou, P., Gkirtzou, K., Deligiannis, M., Galanis, D., Gavriilidou, M., Piperidis, S., Rehm, G., Moritz, M., and Silva, A. G. (2019). Metadata schema, August. ELG Deliverable D2.3. ELG: European Language Grid.

Labropoulou, P., Gkirtzou, K., Gavriilidou, M., Deligiannis, M., Galanis, D., Piperidis, S., Rehm, G., Berger, M., Mapelli, V., Arranz, V., Choukri, K., Backfried, G., Pérez, J. M. G., and Silva, A. G. (2020). Making Metadata Fit for Next Generation Language Technology Platforms: The Metadata Schema of the European Language Grid. Submitted to LREC 2020.

Lösch, A., Mapelli, V., Piperidis, S., Vasiļjevs, A., Smal, L., Declerck, T., Schnur, E., Choukri, K., and Genabith, J. V. (2018). European Language Resource Coordination: Collecting Language Resources for Public Sector Multilingual Information Management. In Nicoletta Calzolari (Conference chair), et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).

Melnika, J., Lagzdiņš, A., Siliņš, U., Skadins, R., and Vasiļjevs, A. (2019a). Requirements and Design Guidelines, June. ELG Deliverable D3.1. ELG: European Language Grid.

Melnika, J., Vasiljevs, A., Skadins, R., and Lagzdins, A. (2019b). User Requirements and Functional Specifications, April. ELG Deliverable D2.1. ELG: European Language Grid.

Moritz, M., Kintzel, F., Elsholz, E., Rehm, G., and Roberts, I. (2019). Base Infrastructure (first release), May. ELG Deliverable D1.2. ELG: European Language Grid.

Piperidis, S., Papageorgiou, H., Spurk, C., Rehm, G., Choukri, K., Hamon, O., Calzolari, N., Gratta, R. D., Magnini, B., and Girardi, C. (2014). META-SHARE: One Year After. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Piperidis, S., Labropoulou, P., Deligiannis, M., and Giagkou, M. (2018). Managing Public Sector Data for Multilingual Applications Development. In Nicoletta Calzolari, et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Piperidis, S., Galanis, D., Deligiannis, M., Gkirtzou, K., Labropoulou, P., Rehm, G., Kintzel, F., Moritz, M., Roberts, I., and Bontcheva, K. (2019). Specification of the ELG platform architecture, June. ELG Deliverable D2.2. ELG: European Language Grid.

Piperidis, S. (2012). The META-SHARE Language Resources Sharing Infrastructure: Principles, Challenges, Solutions. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Rehm, G. and Hegele, S. (2018). Language Technology for Multilingual Europe: An Analysis of a Large-Scale Survey regarding Challenges, Demands, Gaps and Needs. In Nicoletta Calzolari, et al., editors, *Proceedings of the 11th Language Resources and Evaluation Conference (LREC 2018)*, pages 3282–3289, Miyazaki, Japan, 5. European Language Resources Association (ELRA).

Georg Rehm et al., editors. (2012). *META-NET White Pa-*

per Series "Europe's Languages in the Digital Age". Springer, Heidelberg, New York, Dordrecht, London. 31 volumes on 30 European languages. http://www.meta-net.eu/whitepapers.

Georg Rehm et al., editors. (2013). *The META-NET Strategic Research Agenda for Multilingual Europe 2020*. Springer, Heidelberg, New York, Dordrecht, London.

Rehm, G., Uszkoreit, H., Dagan, I., Goetcherian, V., Dogan, M. U., Mermer, C., Váradi, T., Kirchmeier-Andersen, S., Stickel, G., Jones, M. P., Oeter, S., and Gramstad, S. (2014). An Update and Extension of the META-NET Study "Europe's Languages in the Digital Age". In Laurette Pretorius, et al., editors, *Proceedings of the Workshop on Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era (CCURL 2014)*, pages 30–37, Reykjavik, Iceland, 5.

Rehm, G., Hajic, J., van Genabith, J., and Vasiļjevs, A. (2016a). Fostering the Next Generation of European Language Technology: Recent Developments – Emerging Initiatives – Challenges and Opportunities. In Nicoletta Calzolari, et al., editors, *Proceedings of the 10th Language Resources and Evaluation Conference (LREC 2016)*, pages 1586–1592, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Rehm, G., Uszkoreit, H., Ananiadou, S., Bel, N., Bielevičienė, A., Borin, L., Branco, A., Budin, G., Calzolari, N., Daelemans, W., Garabík, R., Grobelnik, M., García-Mateo, C., van Genabith, J., Hajič, J., Hernáez, I., Judge, J., Koeva, S., Krek, S., Krstev, C., Lindén, K., Magnini, B., Mariani, J., McNaught, J., Melero, M., Monachini, M., Moreno, A., Odjik, J., Ogrodniczuk, M., Pęzik, P., Piperidis, S., Przepiórkowski, A., Rögnvaldsson, E., Rosner, M., Pedersen, B. S., Skadiņa, I., Smedt, K. D., Tadić, M., Thompson, P., Tufiş, D., Váradi, T., Vasiļjevs, A., Vider, K., and Zabarskaite, J. (2016b). The Strategic Impact of META-NET on the Regional, National and International Level. *Language Resources and Evaluation*, 50(2):351–374. 10.1007/s10579-015-9333-4.

Rehm, G., Berger, M., Elsholz, E., Hegele, S., Kintzel, F., Marheinecke, K., Piperidis, S., Deligiannis, M., Galanis, D., Gkirtzou, K., Labropoulou, P., Bontcheva, K., Jones, D., Roberts, I., Hajic, J., Hamrlova, J., Kacena, L., Choukri, K., Arranz, V., Mapelli, V., Vasiljevs, A., Anvari, O., Lagzdins, A., Melnika, J., Backfried, G., Dikici, E., Janosik, M., Prinz, K., Prinz, C., Stampler, S., Thomas-Aniola, D., Perez, J. M. G., Silva, A. G., Berrio, C., Germann, U., Renals, S., and Klejch, O. (2020a). European Language Grid: An Overview. Submitted to LREC 2020. Marseille, France.

Rehm, G., Marheinecke, K., Hegele, S., Piperidis, S., Bontcheva, K., Hajič, J., Choukri, K., Vasiļjevs, A., Backfried, G., Prinz, C., Pérez, J. M. G., Meertens, L., Lukowicz, P., van Genabith, J., Lösch, A., Slusallek, P., Irgens, M., Gatellier, P., Köhler, J., Bars, L. L., Aksoriūtė, A., Bel, N., Branco, A., Budin, G., Daelemans, W., Smedt, K. D., Garabik, R., Gavriilidou, M., Gromann, D., Koeva, S., Krek, S., Krstev, C., Lindén, K., Magnini, B., Odijk, J., Ogrodniczuk, M., Ras, E., Rögnvaldsson, E., Rosner, M., Pedersen, B. S., Skadiņa, I., Tadić, M., Tufiş, D., Váradi, T., Vider, K., Way, A., and Yvon, F. (2020b). The European Language Technology Landscape in 2020: Language-Centric and Human-Centric AI for Cross-Cultural Communication in Multilingual Europe. Submitted to LREC 2020. Marseille, France.

Georg Rehm, editor. (2016). *Language as a Data Type and Key Challenge for Big Data. Strategic Research and Innovation Agenda for the Multilingual Digital Single Market. Enabling the Multilingual Digital Single Market through technologies for translating, analysing, processing and curating natural language content*. CRACKER and Cracking the Language Barrier federation, 7. Version 0.9. 04 July 2016. Prepared by the Cracking the Language Barrier federation, supported by the EU-funded projects CRACKER and LT_Observatory.

Georg Rehm, editor. (2017). *Language Technologies for Multilingual Europe: Towards a Human Language Project. Strategic Research and Innovation Agenda*. CRACKER and Cracking the Language Barrier federation, 12. Version 1.0. Unveiled at META-FORUM 2017 in Brussels, Belgium, on November 13/14, 2017. Prepared by the Cracking the Language Barrier federation, supported by the EU-funded project CRACKER.

STOA. (2017). Language equality in the digital age – Towards a Human Language Project. STOA study (PE 598.621), IP/G/STOA/FWC/2013-001/Lot4/C2, March 2017. Carried out by Iclaves SL (Spain) at the request of the Science and Technology Options Assessment (STOA) Panel, managed by the Scientific Foresight Unit (STOA), within the Directorate-General for Parliamentary Research Services (DG EPRS) of the European Parliament, March. http://www.europarl.europa.eu/stoa/.

Vasiljevs, A., Choukri, K., Meertens, L., and Aguzzi, S. (2019). Final study report on CEF Automated Translation value proposition in the context of the European LT market/ecosystem. DOI 10.2759/142151. A study prepared for the European Commission, DG Communications Networks, Content & Technology by Crosslang, Tilde, ELDA, IDC.

Proceedings, pages 137–140
Paris, UNESCO Headquarters, 5-6 December, 2019.

Spoof-Vulnerable Rendering in Khmer Unicode Implementations

Joshua Horton, Ph.D., Makara Sok, Marc Durdin, Rasmey Ty
National Polytechnic Institute of Cambodia
Phnom Penh, Cambodia
joshua_horton@sil.org, makara@keyman.com, marc@keyman.com, rasmeyt2@npic.edu.kh

Abstract

While there are established conventions for typing Khmer using the Unicode Standard, existing systems provide little assistance to users in following the conventions which are thus often ignored. When typing Khmer text, users find that words can be constructed in multiple ways, all of which look 'correct' on-screen. Furthermore, some aspects of Khmer as implemented by common operating systems deviate from the Unicode Standard. This leads to a number of negative outcomes, including phishing and spoofing security risks, poor searchability and complications with natural language processing. This paper identifies issues in the encoding of the Khmer language with the Unicode Standard.

Keywords: Khmer script, text input, Unicode, security, character ordering

Résumé

ទោះបីជាមានគោលការណ៍សម្រាប់ការវាយអក្សរខ្មែរដោយប្រើ "យូនីកូដស្តង់ដារ" ក៏ដោយ ក៏ជំនួយដែលបានផ្តល់ឱ្យអ្នកប្រើប្រាស់ក្នុងការអនុវត្តតាមគោលការណ៍ទាំងនោះ នៅមានតិចតួចនៅឡើយ ។ នេះជាហេតុនាំឱ្យអ្នកប្រើប្រាស់មិនសូវចាប់អារម្មណ៍អើពើ ។ ពេលវាយអត្ថបទជាភាសាខ្មែរ អ្នកប្រើប្រាស់យល់ថាគេអាចវាយពាក្យបានច្រើនរបៀប ហើយវាមើលទៅត្រឹមត្រូវដូចគ្នានៅលើអេក្រង់ ។ ជាងនេះទៅទៀត ផ្នែកខ្លះនៃភាសាខ្មែរដែលត្រូវបានអនុវត្តដោយប្រព័ន្ធប្រតិបត្តិការពេញនិយម មានលក្ខណៈល្បែងពីការកំណត់របស់ "យូនីកូដស្តង់ដារ" ។ បញ្ហានេះនាំឱ្យមានលទ្ធផលអរិជ្ជមានជាច្រើន ដូចជា៖ ហានិភ័យផ្នែកសុវត្ថិភាពដោយសារការឆបោកនិងក្លែងបន្លំ លទ្ធភាពក្នុងការស្វែងរកមានកម្រិតទាប និង ភាពស្មុគស្មាញក្នុងដំណើរការភាសាបែបធម្មជាតិ ។ ការស្រាវជ្រាវនេះរកឱ្យឃើញនូវបញ្ហានានាក្នុងការអ៊ិនកូដនៃភាសាខ្មែរដោយ "យូនីកូដស្តង់ដារ" ។ ឃីមែន (Keyman) មានដំណោះស្រាយចំពោះបញ្ហាទាំងនេះសម្រាប់ភាសាខ្មែរនិងភាសាផ្សេងទៀតផងដែរ ។

1. Introduction

According to *How to Type Khmer Unicode* (Open Forum of Cambodia, 2004), the Khmer (Cambodian) language has 33 consonants, 14 independent vowels, 16 dependent vowels, and 13 diacritics. These are assigned individual code points in the Unicode Standard in the range "U+1780" through "U+17FF". However, given the highly complex structure of the Khmer writing system, Unicode Standard implementers have encountered some ambiguity in how words are constructed from those component codepoints, which could result in vulnerability to spoofing. In this paper, 'Spoof-vulnerable rendering' is used to describe how incorrectly-encoded clusters can be rendered in a manner that could easily be mistaken for correctly-encoded clusters, whether identical pixel-by-pixel or subtly different.

This section illustrates problematic cases this paper seeks to explore. Each example shows a correct encoding, followed by incorrect encodings of the same word rendering identically on common operating systems. The examples show the Unicode codepoints, sample output from Google Chrome 58.0 on Android 6.0.1, and the Google Search results for those encoding.

Unicode codepoints for Khmer characters always begin with "U+17", so only the last two digits will be displayed in our examples.

Case #1: (Subscript + Vowel)

A word with a subscript and a vowel in different orders look exactly alike on-screen. (See Table 1).

(1a)	81	D2	98	C2	9A	ខ្មែរ 29M
(1b)	81	C2	D2	98	9A	ខ្មែរ 175K

Table 1: ខ្មែរ 'Khmer'

Case #2: Subscript + [D2+9A]

Where there are two subscripts in a word and one of them is [D2+9A], either order of the two are rendered identically (See Table 2).

(2a)	9F	D2	8F	D2	9A	B8	ស្ត្រី 4790K
(2b)	9F	D2	9A	D2	8F	B8	ស្ត្រី 471K

Table 2: ស្ត្រី 'woman'

Case #3: Subscript + Consonant Shifter

Error! Reference source not found. shows a consonant shifter placed before a subscript (3b) and after a subscript (3a), producing near identical display.

(3a)	98	D2	99	C9	B6	84	ម្យ៉ាង 452K
(3b)	98	C9	D2	99	B6	84	ម្យ៉ាង 464K

Table 3: ម្យ៉ាង 'one kind'

Case #4: Consonant Shifter + Vowel

Error! Reference source not found. shows a word which could be encoded in four ways, with identical rendering results.

(4a)	9F	CA	B8	ស៊ី	6,160K
(4b)	9F	B8	BB	ស៊ី	117K
(4c)	9F	BB	B8	ស៊ី	129K
(4d)	9F	C9	B8	ស៊ី	7,860

Table 3: ស៊ី 'to eat (for animals, children)'

Case #5: {[BB] or [B6]}+[C6]

The Nikahit sign [C6] will render in a visually similar way when encoded either before or after a dependent vowel (See **Error! Reference source not found.**).

(5a)	80	BB	C6	កុំ	2,860K
(5b)	80	C6	BB	កុំ	4,830
(5c)	85	B6	C6	ចាំ	2,110K
(5d)	85	C6	B6	ចាំ	5,640

Table 4: កុំ '(negation)' and ចាំ 'to wait'

Case #6: [C4] or {[C1]+[B6]} or {[B6]+[C1]}

As shown in Table 6, the example of លោក can be correctly encoded with a single vowel codepoint [C4] but also renders identically when two vowel codepoints are used, even when ordered in reverse.

(6a)	9B	C4	80		លោក 21,900K
(6b)	9B	C1	B6	80	លោក 24K
(6c)	9B	B6	C1	80	លោក 392

Table 5: លោក 'Mr./Sir'

Case #7: [BE] or {[C1]+[B8]} or {[B8]+[C1]}

As with **Case #6**, this composite vowel will render identically whether encoded correctly as a single codepoint, or incorrectly with two codepoints.

(7a)	9F	CA	BE	94		ស៊ើប 619K
(7b)	9F	CA	C1	B8	94	ស៊ើប 3
(7c)	9F	CA	B8	C1	94	ស៊ើប 1

Table 6: ស៊ើប 'to investigate'

Case #8: [D2+8A] and [D2+8F]

In Khmer orthography, these two subscript consonants are the same character. However, in the Unicode Standard, two separate code sequences have been assigned (See **Error! Reference source not found.**).

(8a)	80	8E	D2	8A	B6	9B	កណ្តាល 967K
(8b)	80	8E	D2	8F	B6	9B	កណ្តាល 4,260K

Table 7: កណ្តាល 'Kandal'

Before exploring these cases further, this paper will look at existing research and give a brief introduction on Khmer typography in Unicode.

2. Khmer Typography

Khmer script has been included in the Unicode Standard since the 3rd edition, 1999. In earlier encoding schemes, multiple characters were frequently combined to form what is now a single Unicode character; characters were encoded according to their visual presentation rather than their function. Khmer subscript consonants were assigned individual code-points and many diacritics had multiple codepoints. For users of these encodings, if a word looked correct on-screen, it was considered to be valid.

The Unicode Standard 3.0 took a different route, encoding characters according to their function rather than their presentation. Codepoints could also display differently according to their context in the text store.

In the release of Unicode Standard 4.0 in 2003, the order of consonant shifter and subscripts were reversed. This led to incompatibilities for which a compromise ordering was proposed. *Issues in Khmer Unicode 4.0* (Solá, 2004) suggested allowing a consonant shifter either before or after the subscript: **B {R | {{Z} C}} {S}* {{Z} C} {{Z} V} {O} {ZJ S}**. To date, Solá (2004) has not been incorporated in the Unicode Standard.

Khmer Character Order	**V.**
B {S}* {C} {V} {O}[1]	3.0
B {R | C} {S {R}}* {{Z} V} {O} {S} • B - a base character • R - a robat ([CC]) • C - a consonant shifter • S - a subscript • V - a dependent vowel • Z - a zero width (non-)joiner • O - any other sign • * - (occur more than once)	4.0 5.0 6.0 7.0 8.0 9.0

Table 8: Unicode Standard Versions

3. Data and Methodology

14 words and their corresponding alternatives susceptible to spoof-vulnerable rendering were selected from the top one thousand words in the "Khmer Word Frequency List"[2]. Altogether, 62 items were examined and divided into two groups, valid and invalid text inputs, as defined by conformity to the existing Unicode rules and constraints seen below.

- **Rule #1:** No more than one dependent vowel codepoint is permitted in a syllable (Open Forum of Cambodia, 2004).
- **Rule #2:** A vowel can never be encoded before a subscript (Open Forum of Cambodia, 2004); it can only follow a consonant, a shifter or a Robat sign (Khmer Generation Panel, 2016).

[1] The abbreviations from version 4.0 of the Unicode Standard 4.0 are used here for comparison purposes.

[2] The list is available at: http://sealang.net/project/list/.

- **Rule #3:** When there are two subscripts and one of them is [D2+9A], the other subscript should come first, per spelling order (Open Forum of Cambodia, 2004).
- **Rule #4:** The shifters Triisap[3] [CA] and Muusikatoan[4] [C9] are always typed after the consonant and any subscript but before a vowel.[5] When the shifter is followed by one of the vowels ([B7], [B8], [B9], [BA], [BE], and [B6+C6]), the shifter is rendered as symbol resembling [BB]. Because [BB] itself (Open Forum of Cambodia, 2004) is always a vowel, it should not be used in place of a shifter in this situation, despite the identical appearance.
- **Rule #5:** Robat [CC] must always be inserted after the consonant above which it will be placed, before any vowels. A syllable that has [CC] cannot have any subscript, nor vowels or signs (Open Forum of Cambodia, 2004).
- **Rule #6:** Subscript marker [D2] must occur in between two consonants in a cluster (Khmer Generation Panel, 2016).
- **Rule #7:** When [BB] or [B6] is used with [C6], [C6] must be inserted after [BB] or [B6] (Open Forum of Cambodia, 2004).
- **Rule #8:** Subscripts [D2+8F] and [D2+8A] are usually placed after [93] and [8E] respectively (Sok, 2016). There are some ambiguities as to the identity of the subscript. This needs further discussion.

4. Results and Discussion

The results were rendered in four separate browsers running on Mac OS X 10.12 Sierra and Windows 10 (Safari 10.0, Mozilla Firefox 53.0, Microsoft Edge 38, and Google Chrome 58.0) and four separate platforms (iOS 10, Ubuntu 16, Windows 10 and Android 6.0.1) in order to examine cross-platform consistency.

The ambiguities that arise from different encodings and orderings in these cases leads us to recognize the potential for what we term *spoof-vulnerable rendering*, as it becomes impossible for an individual to visually distinguish whether or not the text is encoded correctly. Thus, there is potential for spoofing online where multiple rendering aliases exist for significant portions of an identifier such as a URL.

For example, the bank ACLEDA អេស៊ីលីដា can be aliased in a manner similar to **Case #4**. A malicious actor could use spoof-vulnerable renderings to deceive users into visiting their fraudulent web property in place of a legitimate site. While similar examples have been found with other scripts (Zheng, 2017), the volume of potential aliases in Khmer magnifies the problem.

Furthermore, spoof-vulnerable rendering increases the complexity of natural language processing (NLP). To be effective, NLP must be preceded by a normalisation phase to remove the redundancy that spoof-vulnerable renderings introduce to the text. This normalisation phase may be difficult to achieve because it must be based on a visual examination of a word. This also results in difficulties with searching content.

While outside the scope of this study, cross-platform inconsistencies are worse with incorrect encodings, hampering document portability.

4.1 Rendering Tests

Since some browsers and/or platforms may have the same text rendering performance, only those which have different rendering performance will be examined in greater detail.

In browsers on Windows and Mac OS X, the results demonstrate that Chrome and Firefox rendered identically: 87% spoof-vulnerable rendering; while Edge and Safari have fewer: 46% and 52% respectively (See Figure 1).

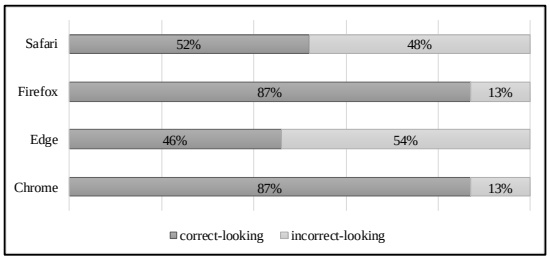


Figure 1: Browser spoof-vulnerable rendering rates

In various platforms, the data shows that iOS achieved 52%, the same as Safari. Ubuntu has the same results as Firefox and Chrome, at 87%. Windows has a result identical to the Edge browser, 46% spoof-vulnerable rendering. Finally, Android surprisingly performs differently from Chrome, suggesting that the rendering systems for Khmer text differ on these two platforms.

In most cases, the valid or correctly-encoded text input is found most frequently online, with the exception of **Case #8,** which had 4,260K results of wrong spelling versus 967K results of correct spelling.

In **Case #2**, the word can possibly be encoded in 8 different ways; and each of them break at least one rule and/or constraint. For instance, (2b) breaks **Rule #3**; (2c) **Rule #8**; (2d) **Rule #3 & #8**; (2e) **Rule #2, #6 & #8**; (2f) **Rule #2 & #6**; (2g) **Rule #2, #3, #6 & #8**; and (2h) breaks **Rule #2**, **#3, & #6.**

[3] A Triisap indicates that vowels accompanying consonants of the first series are to be pronounced as if they accompanied consonants of the second series.

[4] A Muusikatoan indicates that vowels accompanying consonants of the second series should be pronounced as if they accompanied consonants of the first series.

[5] Given the incompatibility between Unicode 3.0 and Unicode 4.0, the character ordering for shifters given precedence by this study is adopted from version 3.0, as (a) it is more consistent with Khmer spelling and pronunciation (Open Forum of Cambodia, 2004), and (b) it is used an order of magnitude more widely in practice.

(2a)	9F	D2	8F	D2	9A	B8	ស្ដ្រី 4,790K
(2b)	9F	D2	9A	D2	8F	B8	ស្ដ្រី 471K
(2c)	9F	D2	8A	D2	9A	B8	ស្ដ្រី 554K
(2d)	9F	D2	9A	D2	8A	B8	ស្ដ្រី 22K
(2e)	9F	D2	8A	B8	D2	9A	ស្ដ្រី 12
(2f)	9F	D2	8F	B8	D2	9A	ស្ដ្រី 3,060
(2g)	9F	D2	9A	B8	D2	8A	ស្ដ្រី 2,470
(2h)	9F	D2	9A	B8	D2	8F	ស្ដ្រី 151K

Table 9: 8 Possible Encoding for ស្ដ្រី

This paper will now examine how **Case #2** is rendered in iOS, Chrome and Edge, and of course Android. Text rendered with unreadable combinations of characters (or dotted circles) is displayed in shaded cells.

	iOS	Chrome	Edge	Android
(2a)	ស្ដ្រី	ស្ដ្រី	ស្ដ្រី	ស្ដ្រី
(2b)	ស្ដ្រី	ស្ដ្រី	ស្ដ្រី	ស្ដ្រី
(2c)	ស្ដ្រី	ស្ដ្រី	ស្ដ្រី	ស្ដ្រី
(2d)	ស្ដ្រី	ស្ដ្រី	ស្ដ្រី	ស្ដ្រី
(2e)	ស្ដ្រ៊	ស្ដ្រី	ស្ដ្រី	ស្ដ្រី
(2f)	ស្ដ្រ៊	ស្ដ្រី	ស្ដ្រី	ស្ដ្រី
(2g)	ស្ដ្រី	ស្ដ្រី	ស្ដ្រី	ស្ដ្រី
(2h)	ស្ដ្រី	ស្ដ្រី	ស្ដ្រី	ស្ដ្រី

Table 10: Rendered Texts of 2a-2h

(2a) show correct encodings for the two words used in this example. (2b) to (2h) show alternate, invalid encodings for those words: they break the rules laid out in **Section 3**. Our data shows that Android is most susceptible to spoof-vulnerable rendering, as all invalid encodings were indistinguishable from the valid encodings. Chrome is also highly susceptible to security attack for rarely does it render text in a way that gives users feedback on the incorrect encoding.

5. Conclusion

This paper has noted several cases in which the Unicode Standard's Khmer script rules can be bypassed while producing visually identical results on many standard browsers and operating systems, which results in spoof-vulnerable renderings. This also has impacts for NLP and document portability.

In order to address spoof-vulnerable rendering, rendering engines in all browsers and platforms should be updated to ensure that invalid encodings do not render identically to valid ones. As the majority of diacritics and vowels in Khmer render with a dotted circle when isolated (to indicate the lack of a base character for combining), we advise that invalid representations be likewise rendered with isolated combining characters to facilitate security and prevent spoofing.

6. Limitations and Future Research

The words were chosen for their potential to demonstrate spoof-vulnerable rendering interactions based on a visual analysis and thus do not represent a comprehensive sample of rendering and encoding issues in Khmer script.

There are a number of other languages that use the Khmer script; these have not been analyzed for this paper and may introduce additional complexities or ambiguities.

While some preliminary rules capable of addressing the example spoof-vulnerable readings of this paper may be quickly identified, development of a more complete set of rules will require further research and testing.

7. Bibliographical References

Open Forum of Cambodia. (2004). *How to Type Khmer Unicode*, Version 1.0:7–14.

Solá, J. (2004). *Issues in Khmer Unicode 4.0. Open Forum of Cambodia*, Version 2.0:6-7.

Khmer Generation Panel. (2016). Association for computing machinery. *Proposal for Khmer Script Root Zone Label Generation Rules*, Version 1.5:15.

Sok, Makara. (2016). *Phonological Principles and Automatic Phonemic and Phonetic Transcription of Khmer Words.* Master's Thesis: 35. Retrieved from: http://inter.payap.ac.th/wp-content/uploads/linguistics_students/Makaras-Thesis.pdf

Zheng, Xudong. (2017). *Phishing with Unicode Domains*. April 14. Online: https://www.xudongz.com/blog/2017/idn-phishing/

Proceedings, pages 141–142
Paris, UNESCO Headquarters, 5-6 December, 2019.

PanLex: A Lexical Infrastructure Tool

David Kamholz, Laura Welcher
The Long Now Foundation, The Long Now Foundation
P.O. Box 475668, San Francisco, CA 94147, USA
kamholz@panlex.org, laura@longnow.org

Abstract

An important piece of language technology infrastructure for under-resourced languages is lexical infrastructure. Word translations have various uses, and for many languages these are among the best data available. The PanLex Database is a lexical infrastructure tool that can translate any word in any language into any other language. It contains 25 million words in 5,700 languages, representing 1.3 billion translations from 2,500 multilingual sources. This practical tool is available now, provides an immediate benefit to linguistic communities, and helps enable future technology. It is an easy, low-cost way to make translation dictionaries available online and interoperable with other languages.

Keywords: translation, dictionary, database

Résumé

[Balinese] Kepahan sané dahat mautama ring kawéntenan infrastruktur teknologi bahasa (panglimbak teknologi basa) ring basa sané nénten makéh maduwé sumber inggih punika infrastruktur leksikal. Artos sajeroning kruna-kruna makéh pisan kawigunanyané, taler ring makudang-kudang basa, data puniki wantah data sané pinih becik kawéntenannyané. Database PanLex inggih punika piranti infrastruktur leksikal sané prasida kaanggén ngartos sakancan kruna ri sajeroning basa sané kaartos dados basa lianan. Puniki madaging slaé (25) yuta kruna ring limang tali pitungatus (5700) basa, pinaka panyeledihi 1,3 miliar artos saking kalih tali limangatus (2500) sumber basa lianan. Piranti praktis puniki mangkin sampun wénten, prasida mawiguna ring paguyuban basa, miwah ngwantu, nyiagayang teknologi riwekas. Puniki dangan pisan, pamargi sané nénten akéh ngamedalang prabea prasida makarya kamus sané wangunnyané online miwah prasida kaoprasiang ring basa lianan.

1. Introduction

Realizing the goal of LT4All requires creating and improving language technology infrastructure for under-resourced languages (98% of all languages). An important piece of this is lexical infrastructure: word translations have a variety of uses, and for many under-resourced languages, lexical data is among the best data available.

The PanLex Database (Kamholz et al. 2014) is a lexical infrastructure tool: it can translate any word in any language into any other language. This practical and useful tool is available now, and can be improved quickly at low cost. It provides an immediate benefit to linguistic communities and helps enable future language technology, such as machine translation.

PanLex (panlex.org) is a project of The Long Now Foundation (longnow.org), a nonprofit that fosters long-term thinking. The PanLex Database has been continuously developed since 2005 as a sister project to the foundation's Rosetta Project which collects parallel documentation on the world's languages for very long-term archiving (see rosettaproject.org).

2. The PanLex Database

The PanLex Database is the world's largest lexical translation database, currently containing 25 million words in 5,700 languages. It has been designed from the start to accommodate lexical data from all languages, including dialects and other subvarieties, without privileging any one language. It has also been designed to accommodate a diverse array of multilingual sources. At minimum, a lexical data source need only provide lexemes in one language and corresponding lexical translations or explanatory definitions in another language. This allows inclusion of basic wordlists and other limited sources. Richer information is included if available, for example part of speech, division into senses, semantic domain, and usage register. The database is available under the CC0 license. Data dumps and a live HTTP API are available (dev.panlex.org).

The PanLex Database incorporates data from more than 2,500 different sources. Sources run the gamut from print to digital, from basic wordlist to large database. The PanLex team particularly emphasizes sources from the least-resourced languages. Each source receives a quality score from 0 to 9.

The PanLex Database can generate both *direct* and *indirect* lexical translations. A translation is direct if one or more sources exist in the database that directly attest the translation of the desired lexical item into the target language. Depending on the language pair and the word being translated, a direct translation may not be available. An indirect translation first translates a word directly into all available languages, producing a set of intermediate translations, and then tries to translate each of these into the target language. Indirect translations are most effective when there are many intermediate translations, as these will effectively converge on the correct target translations. Every available language will be used for intermediate translations; there is no reliance on a single pivot language like English.

Direct and indirect translations are ranked according to translation score, which is assigned based on the number and quality of sources in the database that support the translation. This means that high-quality translations can be

produced from sources of varying quality, and that quality improves over time as new, independent sources are added to the database.

Figure 1 below, taken from PanLex's online translator (translate.panlex.org), illustrates a direct and indirect translation from Breton *dour* 'water' to Cuzco Quechua *yaku*. The double-sided arrow indicates that a direct translation is available between the two words. The other lines show a subset of the available intermediate translations that support an indirect translation; these are in a variety of languages such as French *eau*, Swahili *maji*, and Georgian წყალი. Hundreds of intermediate translations may be used when generating indirect translations.

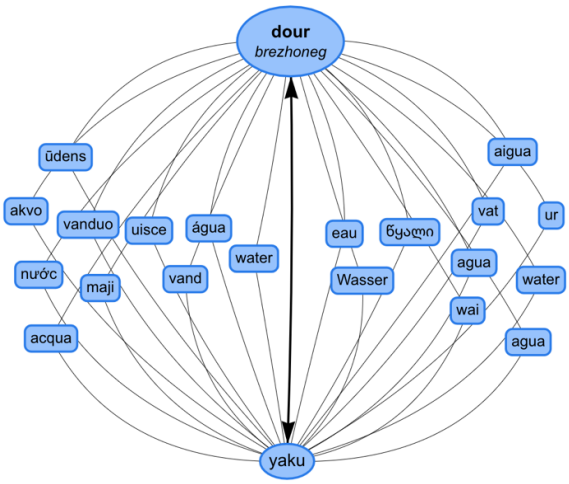


Figure 1: Translation of Breton *dour* 'water' to Cuzco Quechua *yaku*, with subset of intermediate translations.

3. An Enabling Technology

The PanLex Database currently contains 1.3 billion direct translations and billions more indirect translations. It is useful both on its own and as an enabler of other language technology.

Lexical translations have a variety of practical uses. End-users can use the PanLex Database directly in order to gloss words in unfamiliar languages in contexts such as travel, reading, education, and professional translation. It can help create multilingual glossaries for emergency preparedness. The database is an easy, low-cost way to make translation dictionaries available and interoperable with other languages.

Developers can use PanLex data to create their own apps and interfaces. A mobile keyboard app called Polyglot is currently under development which allows language learners to look up and check word translations interactively as they type, in any language for which PanLex has data. PanLex is especially useful where broad language support is needed.

Finally, the PanLex Database helps enable other efforts to bring language technology to under-served languages. It can support localization work by providing seed translations for common terms. It can help improve machine translation quality and coverage as it is developed for more under-resourced languages.

4. Future Steps

PanLex has successfully linked more than 2,500 multilingual sources and made them interoperable, but much work remains to improve lexical infrastructure for under-resourced languages. Many languages need improved coverage, and PanLex already has more than 4,000 sources in its backlog.

The PanLex team is currently in discussions with Wikimedia Deutschland to make PanLex data available through Wikidata (wikidata.org). This will make the PanLex Database available through a well-known international platform and will allow crowd-sourced improvements. We are excited to continue developing this infrastructure to enable linguistic diversity and multilingualism worldwide.

5. Bibliographical References

Kamholz, D., Pool, J., and Colowick, S. (2014) PanLex: Building a Resource for Panlingual Lexical Translation. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 3145-3150, Reykjavik, Iceland. European Language Resources Association (ELRA).

Development of the Parallel Corpus of Mexican Languages (CPLM)

Cynthia Montaño, Gerardo Sierra, Gemma Bel-Enguix
Instituto de Ingeniería
Universidad Nacional Autónoma de México
{cmontanor,gsierram,gbele}@iingen.unam.mx

Abstract

Mexico has a great language diversity. In addition to Spanish, there are 68 language groups and 364 variants (INALI, 2008), divided into 11 families. However, this wealth has been threatened due to discrimination against speakers. Indeed, Spanish has been imposed from the legislative, political and economic point of view, which has interrupted the intergenerational transmission of originary languages and, with it, caused the gradual loss of use spaces and communicative functions. Likewise, few technologies have been developed for these languages, because there are few texts written on the internet. The CPLM is a collaborative parallel corpus that contains texts aligned in Spanish and in six indigenous languages: Mayan, Ch'ol, Mazatec, Mixtec, Otomi and Nahuatl. This article describes the development of the CPLM, as well as the difficulties presented throughout the process.

Keywords: Low-Resources Languages, Parallel Corpus, Indigenous Languages of Mexico

Resumen

México cuenta con una gran diversidad de lenguas, ya que, aparte del español, existen 68 agrupaciones lingüísticas y 364 variantes (INALI, 2008), repartidas en 11 familias. Sin embargo, esta riqueza se ha visto amenazada debido a la discriminación hacia los hablantes. En efecto español se ha impuesto desde el punto de vista legislativo, político y económico, lo que ha interrumpido la transmisión intergeneracional de las lenguas originarias y, con ello, originado la pérdida paulatina de espacios de uso y funciones comunicativas. Así mismo, pocas tecnologías se han desarrollado para estas lenguas, debido a que existen pocos textos escritos en internet. El CPLM es un corpus paralelo colaborativo que presenta textos alineados en español y en seis lenguas indígenas: maya, ch'ol, mazateco, mixteco, otomí y náhuatl. Este artículo describe el desarrollo del CPLM, así como las dificultades presentadas a lo largo del proceso.

Palabras clave: Lenguas de Bajos Recursos, Corpus Paralelo, Lenguas Indígenas de México

1. Introduction

Mexico is one of the most diverse countries linguistically, since it occupies the eighth place worldwide and first in Latin America, followed by Brazil. Despite this, few technological tools have been developed for Mexican languages, which are in danger of extinction, since they have not received the same attention as Spanish, because they have historically been discriminated against. In addition, primary areas for the social welfare of their communities of speakers, such as education and health, have been neglected.

English, French and Spanish, among others, are languages with a large number of speakers, for which numerous linguistic corpus have been built. In contrast, the indigenous languages of Mexico are among the languages of few resources, due to the shortage of written sources to form corpus. To compensate for this, parallel corpus have been constructed in Spanish and in the minority languages of Mexico, since these offer various possibilities that can increase our knowledge about their typological, grammatical and cultural characteristics. In addition, corpora show the differences between genres and their translations.

There are various Natural Language Processing (NLP) tasks that are based on the use of parallel corpora. Some examples are automatic translation, natural language generation, lexical and terminological extraction, morphological segmentation and analysis, part of speech tagging, spelling correction, optical character recognition (OCR), and language identification.

The original languages of Mexico belong to 11 typologically diverse families, each with characteristic features that present particular challenges. Some of the most significant aspects for the treatment of these languages in NLP are the agglutination of morphemes in the Yuto-Nahua family, where Nahuatl is found; the tone in Oto-Mangue languages, which can express both lexical meaning and grammatical function (Suárez, 1973); as well as the ergativity in the Mayan family (Sánchez, 2008). As can be seen, from the perspective of computational linguistics, Mexican languages present a number difficulties.

In general, there is limited production in both digital and printed texts, since in most communities a strong oral tradition is observed, while the written form has not been much encouraged, due to political and social factors that have affected the literacy processes. On the other hand, Mexican languages face a lack of spelling normalization, coupled with great dialect variation, as well as diachronic variation of writing, which represents a challenge in the

processing of these texts when you want to work with NLP.

According to Mager et al (2018), it is important to point out the challenges of working on the development of linguistic resources and tools for the NLP for the languages of Mexico. Addressing these challenges contributes to creating more computational linguistic models, as well as developing a deeper look at the understanding of human language. Additionally, the creation of language technologies in Mexican languages can have a positive social impact on language communities, given the scarcity of digital resources in these languages.

The parallel corpus in Mexican languages that we can find online are Axolotl, a parallel Nahuatl-Spanish corpus, which contains documents of classical and modern Nahuatl (Gutiérrez-Vasques, Sierra and Pompa, 2015) and the Tsunkua project, otomí parallel corpus -español, which contains variants from Mezquital and the State of Mexico. Since these efforts are concentrated in two languages, the UNAM Language Engineering Group proposed to create a parallel corpus that would house several Mexican languages. Thus was born the CPLM.

2. The Parallel Corpus of Mexican Languages (CPLM)

The CPLM is part of an interdisciplinary project whose main objective is to contribute to the development of natural language processing, focused on Mexican languages with limited digital resources -particularly in the task of multilingual lexical extraction- deepening the study of these in terms of models of statistical representation

Among the specific objectives of the project, a methodology for bilingual lexical extraction from parallel corpus of Mexican low-resourced languages is considered. This will allow, for example, to automatically extract bilingual dictionaries and build databases for applications such as machine translation.

Likewise, the project aims to propose one or more types of evaluations that are useful to analyze the effectiveness of the representations and proposed methodology. In addition, we want to explore the development of computational models of various linguistic levels of the treated languages, so that they help in the task of bilingual lexical extraction, mainly morphological segmentation models and syntactic analysis. Finally, it is intended to measure, in quantitative terms, various linguistic phenomena, such as complexity, in order to develop better computational models and contribute from this area to the knowledge and analysis of Mexican languages.

2.1 CPLM Data

The CPLM contains texts in 6 languages belonging to three families: Oto-Manguean, Mayan and Uto-Aztecan. The Oto-Manguean family includes Mixtec, Otomi and Mazatec. Mixtec is spoken in the states of Oaxaca, Guerrero and Puebla and, according to the INALI catalog (2008), presents a total of 81 variants. The Otomi is spoken in the State of Mexico, Hidalgo, Querétaro, Guanajuato, Puebla, Mexico City, Tlaxcala, Veracruz, Michoacán and San Luis Potosí. According to INALI it has 9 variants. Mazatec, spoken in the north of Oaxaca, Puebla and Veracruz, has 16 variants.

Within the Mayan family there are two languages: on the one hand, Ch'ol, which is spoken in the states of Chiapas, Campeche and Tabasco and has two variants: northwest and southeast. On the other hand, the Maya, in the states of Yucatan, Quintana Roo and Campeche. There are some discrepancies regarding the number of Maya variants.

Finally, Nahuatl is the only language of the Yuto-Nahua family present in the corpus. This has 30 variants (INALI, 2008). It spreads through the states of Puebla, Veracruz, San Luis Potosí, Oaxaca, Guerrero, Hidalgo, Colima, Durango, Jalisco, Michoacán, Morelos, Nayarit, Tabasco, Tlaxcala, State of Mexico.

Table 1, shows the languages of the CPLM and the number of variants reported.

Maya	Otomangue	Yuto-nahua
Yucatec Maya (3 variants) Ch'ol (2 variants)	Mazateco (6 variants) Mixteco (30 Variants) Otomí (5 variants)	Nahuatl (5 variants)

Tabla 1: Families, languages and variants

The textual genres that make up the CPLM are: didactic, expository, narrative, poetic, religious, historical and political.

Teaching texts include writing and reading manuals and topics related to language systems. The expository texts include writings of scientific dissemination, for example those dealing with diseases and crops. The stories, traditional fables and of everyday life tales come together in the narrative category. We consider as poetic those texts written in verse. As regards the religious genre, only the Bible is currently available. Historical writings expose the popular history of communities. Finally, the political genre contains articles of the Constitution, as well as explanatory texts on the political-legal field.

Table 2 shows the number of texts for each genre, according to the language.

	Ch'ol	Maya	Mazatec	Mixtec	Nahuatl	Otomí
Didactic	5	5	15	6	5	20
Expositive	7	0	9	12	4	12
Narrative	11	26	28	39	10	66
Poetic	1	5	3	3	11	2
Historic	2	1	1	1	0	1
Polític	2	6	1	5	5	2
Religious	1	1	4	12	10	1

Tabla 2: Genre of the texts

The best represented genre is narrative, since the oral tradition tales are the ones that have been most recorded in the publications of the Summer Linguistic Institute and INALI, the main sources of consultation of the CPLM.
There are three main steps in elaborating this corpus: a) search and compilation of texts, b) digitization and, finally, c) alignment. These steps will be briefly explained in the next section.

3. Elaboration of the corpus

The first step to create the CPLM, was a search of texts published in each of the six languages mentioned above, with their Spanish parallel. Second, the texts were digitized using ABBYY FineReader software, with an OCR that helped prepare the texts for the next stage. Thirdly, the texts were aligned with their corresponding translation in Spanish.

3.1 Text search and compilation

In this step, texts on the internet and libraries were searched. Although the CPLM intends to be a multilingual parallel corpus, in this first stage only texts in the indigenous languages indicated with their respected translation in Spanish were searched. Most of the texts that make up the CPLM were found in PDF format, however, in some cases, bilingual books were found that contained a significant amount of images, for these they were scanned in high definition to facilitate the use of the OCR. In this first phase a database was also created where information on the textual genre, language, variant, ISO code and community was recorded.

3.2 Digitalization

Once all the texts were in PDF, the files were treated with ABBYY FineReader software. This program is used to more easily recognize the common spellings of indigenous languages, such as superscripts, subscripts and diacritics, thanks to OCR character modeling. This saved a significant amount of time during the review of each text and its correct digital transcription.

3.3 Alignment

An important aspect that allows to exploit the bilingual lexical information contained in a parallel corpus is alignment. Alignment is the process of matching bilingual correspondences at a specific level, for example, at the document level, at the paragraph level or at the sentence level and, finally, at the most granular and difficult level to perform, word level alignment.
In general, in CPLM, the texts are aligned at the sentence level. However, in political and religious texts, the alignment is found, either at the constitutional article level or at the verse level.
Initially, the alignment was attempted automatically with the Gale & Church algorithm (1993), used for other parallel corpus. However, since we worked with languages typologically different from those used by Gale & Church, the algorithm was not totally efficient. For example, paragraphs were deleted in Otomi. For this reason, we decided to manually review each of the alignments.

4. The spellings and their difficulties

As already mentioned, there is no general agreement regarding the orthographic norms of the indigenous languages, since there is still a lot of research on the variants that make up the linguistic groups. On the other hand, the written production collected belongs to different years and authors, so generally, the texts are not orthographically homogeneous and present a large number of spellings. The language that shows more orthographic variants is Otomi, since, apart from being a tonal language, it has a large vocal inventory with 9 oral and 5 nasal vowels. Another example in the spelling change is the use of 'h' for the glottal consonant, but its use has been replaced by the apostrophe (').
In Table 3, we present a compilation of the peculiar spellings found in all the CPLM texts.

Spellings
á, à, ǎ, â, ā, a̱, ą, a̰, a̲, á̱, ą́, ą̀, ą̌, ą̂, ą̄, ä, ǟ, æ, ǽ, æ̀, æ̌, æ̂, ǣ, æ̱, æ̰, æ̲, ǽ̱, ǽ̰, æ̰̀, æ̰̌, æ̰̂, ǣ̰
é, è, ě, ê, ē, e̱, ę, ḛ, e̲, é̱, ę́, ę̀, ę̌, ę̂, ę̄, ë, ḗ, ë̀, ë̌, ë̂, ë̱, ɛ, ɛ́, ɛ̀, ɛ̌, ɛ̂, ɛ̄, ɛ̱, ɛ̨, ɛ̰, ɛ̱́, ɛ̨́, ɛ̰̀, ɛ̰̂,
í, ì, ǐ, î, ī, i̱, i̧, ḭ, i̲, í̱, í̧, ì̧, ǐ̧, î̧, ī̧, ï, ḯ, ï̀, ï̱, ï̧, ḭ̀, ḭ́, ḭ̌, ḭ̂, ḭ̄, ḭ̱, ḭ̧, ḭ̲, ḭ̱́
ó, ò, ǒ, ô, ō, o̱, o̧, o̰, o̲, ó̱, ó̧, ò̧, ǒ̧, ô̧, ō̧, ö, ӧ́, ӧ̀, ӧ̱
ú, ù, ŭ, û, ū, u̱, u̧, ṵ, u̲, ú̱, ú̧, ù̧, ǔ̧, û̧, ū̧, ü, ǘ, ǜ, ü̱, ʉ, ʉ́, ʉ̀, ʉ̌, ʉ̂, ʉ̄, ʉ̱, ʉ̰, ʉ̲, ʉ̱́
¢

Tabla 3: Indigenous languages spellings

The spellings in Table 3 represent various linguistic characteristics, whether tone, type of oral vowels different from those of Spanish or nasal vowels.

Each of these involved different challenges. First, during the scanning process, the spellings were not recognized with the OCR, so character molding was used, a special tool of the ABBYY FineReader software. With this tool, each of the unconventional letters were recognized and stored in the software for later use. So, every time a new text was digitized, it recognized well-written spellings and accumulated new ones.

For the alignment, we had to be very careful that the spellings were preserved when passing the texts to the .txt files and especially with the UTF-8 encoding. For that we make sure to find the equivalence of the spellings in ASCII code.

The graphical interface is freely available on the website: http://www.corpus.unam.mx/cplm

5. Conclusion

This article describes the creation of the Parallel Corpus of Mexican Languages (CPLM). The different stages of elaboration have been succinctly presented, as well as the most relevant information. The CPLM was created using GECO, a corpus manager that allows the inclusion of several collaborators, since the CPLM intends to invite students or researchers to participate in the feeding of the CPLM with the corpus that belong to them, either with books or elicitations. With the dissemination of this interface in different forums, we intend to give visibility to Mexican languages in the area of the NLP, in addition to promoting the use of the corpus as a tool to create language technologies.

In future work two lines of work are considered. First, we plan to make improvements to the interface, that is, adaptations will be made so that the CPLM can include recordings and increase the number of texts, as well as add more languages along with their variants. Secondly, it our second goal is to create dictionaries with the vocabulary that many of the texts included in the CPLM contained. Likewise, we will label the texts in Mexican languages in order to perform the search with POS tags.

Regarding the area of the NLP, it is contemplated to work with the analysis and measurement in quantitative terms of the complexity of various linguistic phenomena for each language. The above, in order to understand how to model different types of bilingual relationships depending on the type of languages. Also, another of the future tasks is the creation of bilingual lexical extraction methods based on the distributional vector representations (word embeddings) of word appearance contexts. These models should be able to find word-level correspondences between a pair of languages, based on different statistical approaches of NLP and machine learning techniques. The investigation of these models will be focused on treating typologically distant languages.

9. Acknowledgements

This work is supported by the Mexican Council of Science and Technology (CONACYT) funds A1-S-27780 and FC-2016-01-2225, and PAPIIT IA401219.

6. Bibliographical References

Gale, W. and Church, K. (1993). A program for aligning sentences in bilingual corpora, *Computational Linguistics*, 19:75–102, 01.

Gutiérrez-Vasques, X., Sierra, G., and Hernández, I. (2016). Axolotl: a web accessible parallel corpus for spanish-nahuatl. 05.

Gutiérrez-Vasques, X. (2015). Bilingual lexicon extraction for a distant language pair using a small parallel corpus. pages 154–160, 01.

INALI. (2008*). Catálogo de las Lenguas Indígenas Nacionales: Variantes Lingüísticas de México con sus autodenominaciones y referencias geoestadísticas*

Kann, K., Mager Hois, J. M., Meza-Ruiz, I. V., and Schütze, H. (2018). Fortification of neural morphological segmentation models for polysynthetic minimal-resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long Papers), pages 47–57, New

Orleans, Louisiana, June. Association for Computational Linguistics.

Mager, M., Gutierrez-Vasques, X., Sierra, G., and Meza, I. (2018). Challenges of language technologies for the indigenous languages of the americas. In *Proceedings of the 27th International Conference on Computational Linguistics,* pages 55–69, Santa Fe, New Mexico.

Sierra, G., Solórzano Soto, J., and Curiel Díaz, A. (2017). Geco, un gestor de corpus colaborativo basado en web. *Linguamática*, 9(2):57–72.

Suárez, J. A. (1973). On proto-zapotec phonology. *International Journal of American Linguistics*, 39(4):236 249.

Sánchez, M. E. (2008). Ergatividad en la familia lingüística maya. *Memorias del IV Foro Nacional de Estudios en Lenguas*, 19:541–557, 01.

Proceedings, pages 147–150
Paris, UNESCO Headquarters, 5-6 December, 2019.

Dictionary 4.0: Alternative Presentations for Indonesian Multilingual Dictionaries

Arbi Haza Nasution, Totok Suhardijanto
Informatics Engineering Department Universitas Islam Riau, Linguistics Department Universitas Indonesia
Pekanbaru Riau Indonesia, Jakarta Indonesia
arbi@eng.uir.ac.id, totok.suhardijanto@ui.ac.id

Abstract

Building a multilingual dictionary for 719 languages in Indonesia is a challenging task. We have developed application to create the Leipzig-Jakarta list database for all indigenous languages in Indonesia. The database can be used to generate lexical similarity or lexical distance matrix between languages by comparing the word list. For starter, we covered 11 languages: Indonesian, Javanese, Sundanese, Madurese, Bima, Ternate, Tidore, Palembang Malay, Mandailing Batak, Malay, and Minangkabau. The application has two main features: exploring the existing translations and adding translations to a new language or editing existing translations through crowdsourcing. User acceptance test showed 3.48/4 score.

Keywords: multilingualism, multilingual dictionary, lexical network, lexical computation, computational linguistics

Abstrak

Membangun kamus multibahasa untuk 719 bahasa di Indonesia adalah tugas yang berat. Kami telah mengembangkan aplikasi untuk membuat pangkalan data daftar Leipzig-Jakarta untuk semua bahasa daerah di Indonesia. Pangkalan data tersebut dapat digunakan untuk menghasilkan kesamaan leksikal atau matriks jarak leksikal antar bahasa dengan membandingkan daftar kata tersebut. Sebagai permulaan, aplikasi ini mencakup 11 bahasa: Indonesia, Jawa, Sunda, Madura, Bima, Ternate, Tidore, Melayu Palembang, Batak Mandailing, Melayu, dan Minangkabau. Aplikasi ini memiliki dua fitur utama: menjelajahi terjemahan yang ada dan menambahkan terjemahan ke bahasa baru atau mengedit terjemahan yang ada melalui mekanisme urun daya. Uji keberterimaan pengguna menunjukkan skor 3,48 / 4.

1. Introduction

According to (Eberhard et al., 2019), there are 719 languages in Indonesia, where 707 languages are still alive and 12 languages have become extinct. Extinct in this sense is that there are no longer any of the speakers. Among the surviving languages, 701 languages are local languages and 6 languages are not local languages. Furthermore, there are 18 languages that are used as administrative and / or educational languages, 73 languages are still growing, 188 languages are classified as strong, 347 languages are in difficulty, and 81 languages are in a danger of extinction.

Furthermore, based on his observations, (Anderbeck, 2015) groups Indonesian languages into three groups. First, about two of the four languages in Indonesia today still have a vital life force and have a safe number of speakers (EGIDS (Expanded Graded Intergenerational Disruption Scale) 1-6a). In this group, intergenerational transmission of speakers still occurs and persists. Even though some of them are bilingual, they know when to use local and Indonesian languages. Second, one of the four languages in Indonesia is in fragile condition (EGIDS 6b Threatened) with speakers who continue to decline in number. Usually most young people still learn their mother tongue, but certain reasons make them change their orientation towards languages that are more economically advantageous. Third, the rest, one of the four languages in Indonesia seems to be dying (EGIDS 7-8b) or may have become completely extinct (EGIDS 9 and 10). Some, like the Marori language, may be lost in a generation. The other may be in two or three generations. With conditions like that, of course, we are like racing with time to document language.

Although some experts distinguish the terms of language documentation and language description (Austin and Sallabank, 2011), in some ways, the two are interconnected. According to Austin, the documentation and description of languages differ in their purpose, points of interest, research methods, workflow, and outcomes. Descriptions or language descriptions basically aim at producing grammar, dictionaries, and collections of texts, the target users are generally linguists, and the material produced is sometimes written in a framework that is accessible to trained linguists. In contrast, language documentation is discourse-centered, the main objective being the direct representation of as many types of discourse as possible (Austin, 2007; Woodbury, 2003; Himmelmann, 1998). However, according to (Austin and Grenoble, 2007) the documentation project must rely on the application of theoretical and descriptive linguistic techniques so that the resulting output is sure to be utilized and understood by many communities. So, in other words, documentation and description are activities with objectives and outcomes that complement each other, and one of their important outcomes is the result of lexicographic work, the dictionary.

In the context of endangered languages, dictionaries have a very crucial role, namely storing what is left of endangered languages and cultures by recording valuable information that might be lost (Cristinoi and Nemo, 2013). The bilingual dictionaries are also useful for natural language processing researchers, especially for those related with enrichment of language resources like bilingual dictionary (Nasution et al., 2016; Nasution et al., 2017b; Nasution et al., 2017a; Nasution et al., 2018) or machine translation

(Nasution et al., 2017c; Nasution, 2018). Furthermore, in many cases, the existence of a dictionary can help revive a language and change the attitudes of speakers of that language which ultimately encourage them to use it as often as possible. Even so, (Cristinoi and Nemo, 2013) mentioned that there are some problems related to lexicography in the realm of language documentation. First, the compilers of endangered language dictionaries are generally people or linguists who care. Certainly, the result is different from the general dictionary compiled by a professional team. Secondly, dictionaries made for endangered languages are certainly far from direct economic profit. Third, the endangered language dictionaries have limited distribution, that is only to linguists or the public who have an interest in the language concerned. Fourth, in the work of lexicography in endangered languages there are several problems that must be resolved, for example what characters are used, which variations are considered standard, and so on. Fifth, data collection of endangered languages is more difficult because it only relies on the ethnographic work of researchers or notes from concerned community members. Sixth, the dictionary of threatened languages ??is usually used for research purposes, documenting specific languages and cultures, protecting language and cultural heritage that will be lost without written traditions on the language or culture, helping indigenous people communicate in dominant foreign languages, helping non-native speakers to understand the native speakers and their cultural background, and provide orthography or standard written form for the entire vocabulary.

Because of the problems mentioned above, the data collection of endangered language dictionaries is generally done with a limited number of vocabularies, generally focus on general vocabulary or even basic vocabulary lists. The list is a lexical artifact which is a vocabulary whose references are universally available in many languages in the same region. In the condition of Indonesia which is multilingual, of course the problem becomes more complex. Over time, how do lexicographic studies contribute to language documentation efforts, especially in terms of recording important and varied information about language and culture in Indonesia? Making multilingual dictionaries is not an easy task, especially from the point of computational lexicography (Walker, 1995). Thus, in this paper, we try to build a model that can accommodate the diversity of languages in Indonesia. This can be further elaborated with the question: how to compile dictionaries for languages in Indonesia? What is the correct format of multilingual dictionaries that can help document languages in Indonesia? These two questions will be answered in this paper.

2. Methodology

In the 1950s, linguist Morris Swadesh published a list of 200 words called the Swadesh list, which were thought to be 200 lexical concepts found in all languages ??that were most unlikely to be borrowed from other languages (Swadesh, 1955). Swadesh then reduced the list to 100 items based on intuition where a drastic removal from a 200-word list was the best solution, with the consideration that quality is at least as important as quantity. Although the new list has weaknesses, but the list is relatively light to process because of the small amount. Automated Similarity Judgment Program (ASJP) (Brown et al., 2008) is an open source software with the main objective to develop a Swadesh list database for all languages in the world where lexical similarity or lexical distance matrix between languages can be obtained by comparing the word list. However, the list of 100 Swadesh words was cut down to 40 words that are considered the most stable of forms of change, maintained over time and not replaced by other lexical items from the language itself or elements borrowed from other languages (Holman et al., 2008). The lexical distance between regional languages in Indonesia has been visualized using the ASJP database (Nasution and Murakami, 2019; Nasution et al., 2019). However, there are doubts about the validity of the lexical distance between some regional languages such as between Sundanese and Javanese which should be closer to the lexical distance but only 21.8% of lexical similarities are produced. Therefore, alternative word lists are needed that can produce more accurate lexical distances.

In addition to the Swadesh list, linguists also use the Leipzig-Jakarta list (100 words) (Tadmor et al., 2010) to test the level of chronological separation of languages by comparing words that are resistant to loans. The Leipzig-Jakarta list is available in 2009 (Sakel and Everett, 2012). The mobile application developed in this paper aims to develop the Leipzig-Jakarta list database for all regional languages in Indonesia where lexical similarity or lexical distance matrix between languages can also be further obtained by comparing the word list. The application built will be tested for user satisfaction with quantitative analysis using a questionnaire. The proposed framework is depicted in Figure 1. The data will be used to generate visualization of Indonesian Indigenous Languages Lexical Similarity with Knowledge Graph.

3. Results

For the initial research, 100 Leipzig-Jakarta word lists were translated into 11 languages: Indonesian, Javanese, Sundanese, Madurese, Bima, Ternate, Tidore, Palembang Malay, Mandailing Batak, Malay, and Minangkabau. The application has two main features: exploring the translations of the 100 Leipzig-Jakarta word list and adding translations to new languages or changing translations that are already available. The exploration interface for translating 100 Leipzig word lists into 11 languages with the details of the translated words including the definition, synonyms and example use of the word in a sentence are shown in Figure 2.

To add a translation to a new language or change an already available translation, the user should register to the system first using the registration form. After entering the user's email address for verification, the user can click on the language selection dropdown, then the user can choose the destination language according to the language selection feature, the last step, the user can type the translation according to the language of choice, and click the "*SUNTING / TAMBAH KATA*" (which translated to EDIT / ADD WORDS) button, then the translation added / edited will be

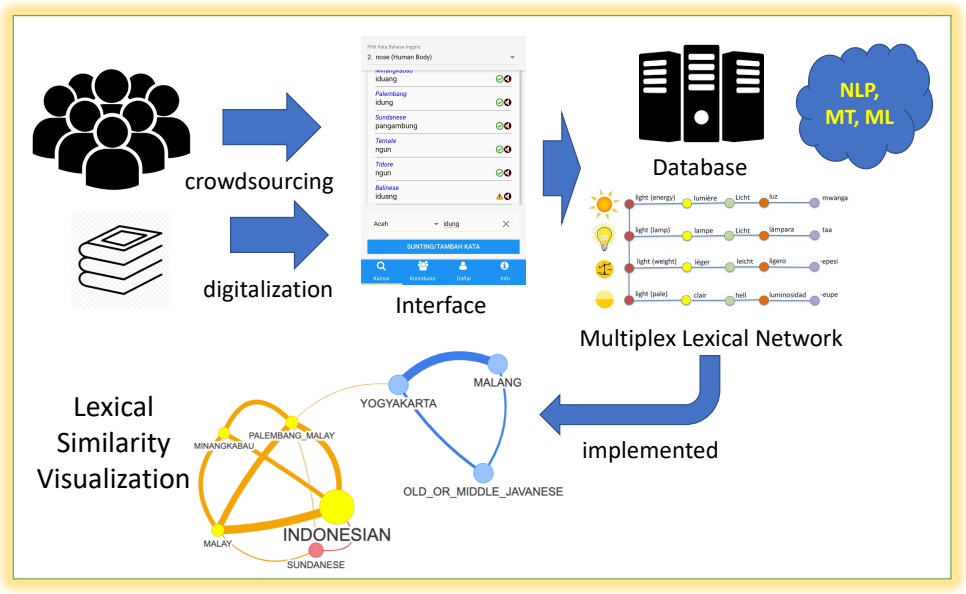


Figure 1: Proposed framework.

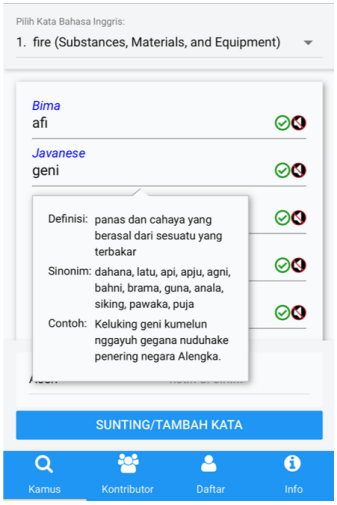


Figure 2: The definition, synonyms and example in sentence.

Figure 3: Leader board of contributor.

verified by the linguist. Finally, the user will get a poin for each translation added or edited, and another poin when the new translation or edition has been verified. The leaderboard is shown in Figure 3.

The application that was built was tested by 36 random users with quantitative analysis using a questionnaire with 7 questions as shown in Table 1. Based on the results of the user satisfaction questionnaire with dictionary 4.0, the average value for the whole questionnaire item was 3.48. This shows that the design and appearance of the Dictionary 4.0 Application is quite interesting, easy to use and accepted by users.

4. Conclusion

Until now, in this study, a multilingual dictionary prototype model with the functionality to collect data of various languages was quickly compiled. Therefore, the focus in this paper is on the issue of setting up a language data collection system through a crowd sourcing mechanism. Meanwhile, in terms of usage, acceptance testing has been carried out to see how well the application design according to the user. Based on these tests, we obtained quite interesting results, which is 3.48 from a scale of 4. The next stage of this research is to upgrade the dictionary 4.0 application that is capable of managing multilingual dictionary services with dedicated functions for general users and registered users. In addition, the language similarity comparison function

Item	Mean	Median	Standard Deviation
Appealing design and appearance	3.47	3	1.078
The design and appearance of the application is easy to understand	3.41	3	1.043
The navigation menu is easy to understand	3.37	3.5	1.157
The colors used in the application are suitable and not excessive	3.72	4	0.958
The application is easy to use	3.47	4	1.078
Easy to explore each word translation	3.47	4	1.047
It is easy to propose revision to existing translations or add new translations	3.47	3.5	1.047

Table 1: Results of user satisfaction questionnaire of dictionary 4.0

will be included using the lexical distance approach as in the ASJP database program.

5. Acknowledgements

This research was partially supported by Universitas Islam Riau.

6. Bibliographical References

Anderbeck, K. (2015). Portraits of language vitality in the languages of indonesia. *Language documentation and cultural practices in the Austronesian world: Papers from*, pages 19–47.

Austin, P. K. and Grenoble, L. (2007). Current trends in language documentation. *Language documentation and description*, 4:12–25.

Austin, P. K. and Sallabank, J. (2011). *The Cambridge handbook of endangered languages*. Cambridge University Press.

Austin, P. K. (2007). Training for language documentation: Experiences at the school of oriental and african studies. *Documenting and revitalizing Austronesian languages*, pages 25–41.

Brown, C. H., Holman, E. W., Wichmann, S., and Velupillai, V. (2008). Automated classification of the world? s languages: a description of the method and preliminary results. *STUF-Language Typology and Universals Sprachtypologie und Universalienforschung*, 61(4):285–308.

Cristinoi, A. and Nemo, F. (2013). Challenges in endangered language lexicography.

Eberhard, D. M., Simons, G. F., and Fennig, C. D. (2019). Ethnologue: Languages of the world.

Himmelmann, N. P. (1998). Documentary and descriptive linguistics.

Holman, E. W., Wichmann, S., Brown, C. H., Velupillai, V., Müller, A., and Bakker, D. (2008). Explorations in automated language classification. *Folia Linguistica*, 42(3-4):331–354.

Nasution, A. H. and Murakami, Y. (2019). Visualizing language lexical similarity clusters: A case study of indonesian ethnic languages. *Journal of Data Science and Its Applications*, 2(2):45–59.

Nasution, A. H., Murakami, Y., and Ishida, T. (2016). Constraint-based bilingual lexicon induction for closely related languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3291–3298, Paris, France, May.

Nasution, A. H., Murakami, Y., and Ishida, T. (2017a). A generalized constraint approach to bilingual dictionary induction for low-resource language families. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 17(2):9:1–9:29, November.

Nasution, A. H., Murakami, Y., and Ishida, T. (2017b). Plan optimization for creating bilingual dictionaries of low-resource languages. In *2017 International Conference on Culture and Computing (Culture and Computing)*, pages 35–41, Sept.

Nasution, A. H., Syafitri, N., Setiawan, P. R., and Suryani, D. (2017c). Pivot-based hybrid machine translation to support multilingual communication. In *2017 International Conference on Culture and Computing (Culture and Computing)*, pages 147–148, Sept.

Nasution, A. H., Murakami, Y., and Ishida, T. (2018). Designing a collaborative process to create bilingual dictionaries of indonesian ethnic languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3397–3404, Paris, France, may. European Language Resources Association (ELRA).

Nasution, A. H., Murakami, Y., and Ishida, T. (2019). Generating similarity cluster of indonesian languages with semi-supervised clustering. *International Journal of Electrical and Computer Engineering (IJECE)*, 9(1):1–8.

Nasution, A. H. (2018). Pivot-based hybrid machine translation to support multilingual communication for closely related languages. *World Transactions on Engineering and Technology Education*, 16(2):12–17.

Sakel, J. and Everett, D. L. (2012). *Linguistic fieldwork: A student guide*. Cambridge University Press.

Swadesh, M. (1955). Towards greater accuracy in lexicostatistic dating. *International journal of American linguistics*, 21(2):121–137.

Tadmor, U., Haspelmath, M., and Taylor, B. (2010). Borrowability and the notion of basic vocabulary. *Diachronica*, 27(2):226–246.

Walker, Z. C. (1995). *Automating the lexicon: research and practice in a multilingual environment*. Oxford University Press.

Woodbury, A. C. (2003). Defining documentary linguistics. *Language documentation and description*, 1(1):35–51.

Proceedings, pages 151–154
Paris, UNESCO Headquarters, 5-6 December, 2019.

MultiTAL
An online Platform to List NLP Tools for Under-Resourced Languages

Damien Nouvel, Mathieu Valette, Driss Sadoun
ERTIM
2, rue de Lille, 75007 PARIS
{damien.nouvel, mvalette}@inalco.fr, driss.sadoun@postlab.fr

Abstract

The diversity and variety of human languages raises indisputable difficulties for processing textual data. Regarding under-resourced languages, many software solutions have been designed, but many are poorly referenced and documented. The ERTIM (INALCO) lab published in 2016 a platform named MultiTAL that addresses this issue. Our platform lists tools available for languages. For each software, the knowledge base provides information concerning : processing tasks, implemented method, OS compatibility, among others. We do not claim to be comprehensive, but people populating the knowledge base are speakers of concerned languages, they downloaded and tested softwares, and provided detailed technical information for installation and use.

Keywords: Multilinguality, NLP Tools

Résumé

La diversité et la variété des langues humaines donne d'incontestables difficultés pour le traitement de données textuelles. Concernant les langages peu dotées, de nombreux logiciels ont été implémentés, mais beaucoup restent peu référencés et mal documentés. L'équipe ERTIM a mis en ligne en 2016 la plateforme MultiTAL qui réalise ce travail. Cette base de connaissances apporte des informations sur des outils, par langue et tâche. Nous ne prétendons pas être exhaustifs, mais les personnes remplissant la base étaient locuteurs des langues concernées, elles ont téléchargé et testé les outils, et ont renseigné des informations détaillées sur leur installation et leur utilisation.

1. Introduction

1.1. A Project for Under-Resourced Languages

It is a known issue that with 150 written languages, a great number of them are considered "under-resourced" from a Natural Language Processing (NLP) point of view. This is indeed even more true for the largest number[1] of oral languages, but we won't consider them in the present work. In the context of globalisation and digitalization, this concern is even more serious as language communities require to access the information-based society and the Internet for various purposes of their everyday life.

Unsurprisingly, languages that are the best equipped with digitalized linguistic resources (e.g. corpora, lexicons, software) are those that have either a long history related to computers or a sufficient economic weight to receive more recent developments. On the other side, those that are not in this case are most of the time very late on such developments. The related communities are often forced to use one of the *lingua franca* already well established on the Internet. As a side-effect, this also raises the risk of language impoverishment. Yet, making those languages exist in the digital world is undoubtedly a necessary step and can't be avoided. When no economic benefits are in sight, those development can only be handled by local or international non-profit organisations such as governments, NGOs, associations or academics.

Our institute, INALCO (National Institute for Oriental Languages and Civilizations) is both a university that teaches around 100 oriental languages from Central Europe, Africa, Asia, America and Oceania and a research center that works on related languages. Within this institute, our research team, ERTIM, frequently collaborates with researchers and teachers on a number of those languages. This position requires that we can quickly establish what is the status of a language in terms of digitalized resources.

To help us with this, we initiated the *MultiTAL*[2] project in 2016, hosted at (`http://multital.inalco.fr`. The *MultiTal* infrastructure aims at providing systemic descriptions of tools (software) for under-resourced languages, so as to document them, promote them, and ease their accessibility. For this purpose, tools are downloaded, installed and tested, we select those that are actually operational, and provide accurate and *critical* documentation, rather than providing lists of tools that have not been tested except by their designers. We also ranked this project as "highly multilingual": the main pages of our website have been localized into 7 languages, so that the largest number of people can understand it, even if they have limited knowledge of English.

1.2. Related Work

Over the last decade, the number of digitized materials has considerably grown. The willingness to take into account this new digital content has led to the popularization of the use of Language Resources (LR) and NLP technologies. However, LR are still difficult to find because they are drowned in the mass of web content. Moreover, their documentation is often monolingual and written either in the developers' languages (such as Arabic, Chinese, Japanese or

[1] Depending on studies, 6000 or 7000

[2] TAL stands for *Traitement Automatique des langues*, as "Natural Language Processing" in French

Russian) or in a *lingua franca* (such as English or French). This situation makes it difficult for scholars to use or re-use LR that could be useful for their work or research. Hence, storing and distributing LR has become an issue in itself. This has been addressed by many initiatives all around the world (Váradi et al., 2008; Tohyama et al., 2008; Piperidis, 2012; Tonne et al., 2013; Calzolari et al., 2012). These initiatives are essential to promote the research and development of language technologies. They also may provide a real picture of tools and resources that are currently available for several languages (Skadina et al., 2013; TADIĆ, 2012; Del Gratta et al., 2014).

In order to describe and share LR, different meta-data models have been proposed (Gavrilidou et al., 2011; Broeder et al., 2012; McCrae et al., 2015a). The models of each provider depend on their coverage and the kind of LRs they manage. Hence, there are as many meta-data models for describing LRs, which may represent a limit for resource sharing and lead to the re-creation of already existing LR resources (Cieri et al., 2010). To address this issue, different attempts have been made, such as an initiative for harmonising between ELRA and LDC catalogs (Cieri et al., 2010) and more recently ontologies were used to devise interconnections among resources (Chiarcos, 2012) or to make meta-data available from different sources under a common scheme (McCrae et al., 2015a; McCrae et al., 2015b). In the perspective of a possible interoperability between our meta-data model and the existing ones, we chose to use an ontology for storing *MultiTal* infrastructure data.

Most existing LR infrastructures focus on occidental languages and invite developers of resources or tools to describe them themselves. Even if it eases access to LR technologies, when it concerns NLP tools it does not necessarily make their use any easier. Often, their usage instructions remain poorly documented. In our project, we aim to list NLP tools processing written non-occidental languages or more precisely languages taught at INALCO. In this framework, each NLP tool is identified, tested and fully documented by an intern speaking the language the tool processes. Then, if the tool appears to run correctly its information is stored within our meta data model (ontology) and its resulting documentation is made available. Our aim is to ensure that tools described on *MultiTal* infrastructure can be properly installed and executed by end-users. As MultiTal's end-users may not be language technology experts and their mother tongue may vary, we use an ontology verbalisation method (Androutsopoulos et al., 2014; Cojocaru and Trãuşan Matu, 2015; Keet and Khumalo, 2016) detailed in (Sadoun et al., 2016) to automatically produce documentation in multiple languages. So that we provide end-users with simple, structured and organised documents containing NLP tool information and detailing instructions of how to install, configure and run tools fitting their needs.

2. The MultiTAL Platform

2.1. Goals

As already stated in Section 1.1., issues are identified and well-known for many languages. Among those, we focus on three main difficulties:

- Relevant NLP tools are not so easy to find, in particular for under-resourced languages
- Documentation is not always comprehensive (sometimes native)
- Instructions to install, configure and execute NLP tools are not that simple

Facing those issues, our goal is to make technologies more accessible regardless of the expertise or spoken language of end-users by means of providing information with a simple, multilingual, structured and standardised tool documentation.

2.2. Methodology

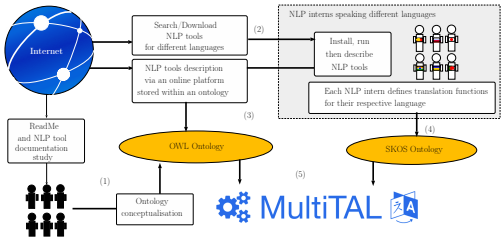


Figure 1: Process of the MultiTAL process

As depicted in 1, we rely mainly on human contributors to enrich our documentation of NLP tools. Their first objective is to verify that the software can easily be downloaded, installed and executed on a standard and minimally configured system (whether Windows, Linux or MacOS) and that the output has minimal accuracy regarding the concerned task so that it can be used out-of-the-box. We believe this to be a major advantage of our platform: we guarantee that we have succeeded in those preliminary steps.

Here is the list of information we do provide about tools:

- The name
- Last update (at the moment we documented it)
- Main programming language
- Accessibility (download, online or web service)
- Licenses
- Authors
- References (mostly academics)
- Description
- Implemented "tasks" with for each:
  - Type(s) of task(s)
  - NLP methods
  - Languages covered
  - Input and output (types and encodings)
- Installation and execution procedures (see below)

We did extensive documentation work for the last item regarding installation and execution. The documentation also describes and gives concrete examples of commands that may be issued,as depicted in Figure 2, for the MeCAB[3] tool. Again, we are driven by end-users consideration, especially those people working with NLP tools that are not necessarily developers but may have a minimum knowledge of using computers to execute this kind of tools by using the command line. In this documentation process, validation is important: tools are entered into the knowledge base, but are not published until the contributor has actually tested the tool.

Installation procedure [Linux]

1 - download mecab-0.996.tar.gz from http://taku910.github.io/mecab/#download

```
download mecab-0.996.tar.gz http://taku910.github.io/mecab/#download
```

2 - uncompress -xzvf mecab-0.996.tar.gz

```
tar -xzvf mecab-0.996.tar.gz
```

3 - go to directory mecab-0.996

```
cd mecab-0.996
```

4 - type the command: ./configure

```
./configure
```

5 - type the command: make

```
make
```

6 - type the command: sudo make install

```
sudo make install
```

Execution procedure [Linux]

1 - type the command: mecab input_file.txt -O output_format -o output_file.txt

```
mecab  input_file.txt -O output_format -o output_file.txt
```

Figure 2: Installation and Execution Documentation

2.3. Architecture

Behind the scenes, our platform is a simple LAMP[4] webserver that is used both for public access and for backend administration of the knowledge base. The latter is an OWL ontology where each tool is described using the properties as reported in Figure 3. Our conceptualisation of NLP processing raised a number of questions: we have been led to distinguish a tool from the tasks it accomplishes (each task is thus validated separately upon testing). In practice, we observed that indeed a tool may have very different accuracy for a given task depending on the language, but we didn't enter into this level of detail.

The platform is hosted at `http://multital.inalco.fr`. By default, the search engine displays a view on tasks, for the sake of users looking for a tool that implements a specific type of processing on a particular language. User interface has been more or less extensively translated into 7 languages : Arabic, Chinese, English, French, Hindi, Japanese, Russian, Tibetan. Both the ontology and the SKOS can be queried using SPARQL or downloaded as RDF/XML files.

[3] `http://taku910.github.io/mecab/`

[4] Linux Apache MySQL PhP

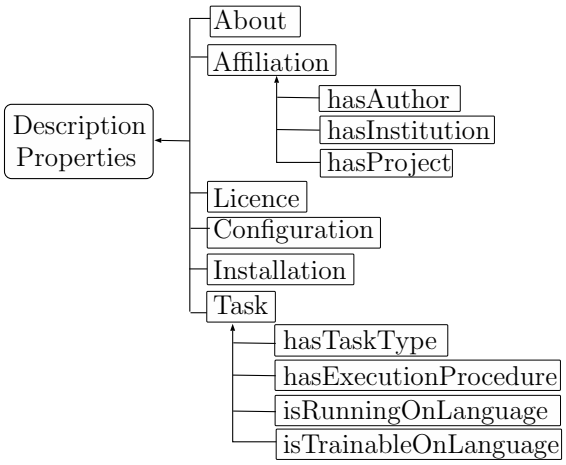


Figure 3: Ontology Properties

2.4. Current status

Currently, the platform documents:

- 45 NLP task types
- 47 languages
- 91 of them are published (among 167 tested)
- 33 NLP methods
- 112 tool authors

Tables 1 and 2 give a rough overview of the documentation per task and language. Be aware that displayed numbers contain redundancy, since a tool may actually implement multiple processings (as tasks) for each language.

Task Type	# Tasks
Part-of-speech tagging	36
Segmentation	32
Tokenization	18
Morphological analysis	15
Lemmatisation	13
Morphological tagging	13
Parsing	12
Dependency parsing	9
Named Entity Recognition	8
Stemming	8
Transcription	6
Transliteration	6
Concordances	5
Diacritization/vocalization	5

Table 1: Tasks by Tasks types

3. Conclusion

In this paper, we described our our MultiTAL platform, which provides documentation for NLP tools restricted to some of the languages taught in our institute INALCO

Language	# Tasks
Arabic	47
Chinese - Mandarin (simplified)	35
Chinese - Mandarin (traditional)	33
Japanese	22
Russian	16
Hindi	15
English	12
French	9
Bulgarian	5
Hungarian	5
Tibetan	5
Ukrainian	5

Table 2: Tasks by Language

(most of the time, oriental and under-resourced languages). Our platform not only provides a structured description of the tool, but also includes additional features. First, it guarantees that we were able to use the tool (by downloading and executing it or online). Second, we detail documentation on the aspect of the installation and execution procedures as commented scripts. The user interface has been translated in 7 languages. We believe our project may be a starting point to establish guidelines and best practices for improving NLP tool documentation.

4. Acknowledgements

Many thanks to all our contributors! This work was founded by the USPC (université Sorbonne-Paris-Cité).

5. Bibliographical References

Androutsopoulos, I., Lampouras, G., and Galanis, D. (2014). Generating natural language descriptions from OWL ontologies: the naturalowl system. *CoRR*, abs/1405.6164.

Broeder, D., Van Uytvanck, D., Gavrilidou, M., Trippel, T., and Windhouwer, M. (2012). Standardizing a component metadata infrastructure. In *the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 1387–1390. European Language Resources Association (ELRA).

Calzolari, N., Gratta, R. D., Francopoulo, G., Mariani, J., Rubino, F., Russo, I., and Soria, C. (2012). The LRE map. Harmonising Community Descriptions of Resources. In *LREC*, pages 1084–1089.

Chiarcos, C. (2012). Ontologies of linguistic annotation: Survey and perspectives. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*.

Cieri, C., Choukri, K., Calzolari, N., Langendoen, D. T., Leveling, J., Palmer, M., Ide, N., and Pustejovsky, J. (2010). A road map for interoperable language resource metadata.

Cojocaru, D. s. A. and Trãuşan Matu, c. (2015). Text generation starting from an ontology. In *Proceedings of the Romanian National Human-Computer Interaction Conference - RoCHI*, pages 55–60.

Del Gratta, R., Frontini, F., Khan, A. F., Mariani, J., and Soria, C. (2014). The lremap for under-resourced languages. *CCURL 2014: Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era*, page 78.

Gavrilidou, M., Labropoulou, P., Piperidis, S., Francopoulo, G., Monachini, M., Frontini, F., Arranz, V., and Mapelli, V. (2011). A metadata schema for the description of language resources (lrs). *Language Resources, Technology and Services in the Sharing Paradigm*, page 84.

Keet, C. M. and Khumalo, L. (2016). Toward a knowledge-to-text controlled natural language of isizulu. *Language Resources and Evaluation*, pages 1–27.

McCrae, J. P., Labropoulou, P., Gracia, J., Villegas, M., Rodríguez-Doncel, V., and Cimiano, P., (2015a). *The Semantic Web: ESWC 2015 Satellite Events*, chapter One Ontology to Bind Them All: The META-SHARE OWL Ontology for the Interoperability of Linguistic Datasets on the Web, pages 271–282.

McCrae, J., Cimiano, P., Doncel, V., Vila-Suero, D., Gracia, J., Matteis, L., Navigli, R., Abele, A., Vulcu, G., and Buitelaar, P. (2015b). Reconciling Heterogeneous Descriptions of Language Resources. *ACL-IJCNLP 2015*, page 39.

Piperidis, S. (2012). The meta-share language resources sharing infrastructure: Principles, challenges, solutions. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*.

Sadoun, D., Mkhitaryan, S., Nouvel, D., and Valette, M. (2016). Readme generation from an owl ontology describing nlp tools. In *2nd International Workshop on Natural Language Generation and the Semantic Web*.

Skadina, I., Vasiljevs, A., Borin, L., LindÃ©n, K., Losnegaard, G., Pedersen, B. S., Rozis, R., and De Smedt, K. (2013). Baltic and nordic parts of the european linguistic infrastructure. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, pages 195–211.

TADIĆ, T. V. M. (2012). Central and south-east european resources in meta-share. In *24th International Conference on Computational Linguistics*, page 431.

Tohyama, H., Kozawa, S., Uchimoto, K., Matsubara, S., and Isahara, H. (2008). Construction of a metadata database for efficient development and use of language resources.

Tonne, D., Rybicki, J., Funk, S., and Gietz, P. (2013). Access to the dariah bit preservation service for humanities research data. In *Parallel, Distributed and Network-Based Processing (PDP), 21st Euromicro International Conference*, pages 9–15.

Váradi, T., Wittenburg, P., Krauwer, S., Wynne, M., and Koskenniemi, K. (2008). Clarin: Common language resources and technology infrastructure. In *6th International Conference on Language Resources and Evaluation (LREC 2008)*.

Proceedings, pages 155–159
Paris, UNESCO Headquarters, 5-6 December, 2019.

A Speaking Atlas of Indigenous Languages of France and its Overseas

Philippe Boula de Mareüil[1], Frédéric Vernier[1], Gilles Adda[1], Albert Rilliard[1], lacques Vernaudon[2]

[1]LIMSI, CNRS & Université Paris-Saclay, Orsay, France
[2] EASTCO, Université de la Polynésie française, Faaa, PF
{philippe.boula.de.mareuil, frederic.vernier, gilles.adda, albert.rilliard}@limsi.fr
jacques.vernaudon@upf.pf

Abstract

The objective of this work is to show and valorise the linguistic diversity of France through recordings collected in the field, a computer interface (which allows viewing the dialect areas) and a work of orthographic transcription. We briefly describe a website (https://atlas.limsi.fr) presenting interactive maps of France and its overseas territories, from which the Aesop fable "The North Wind and the Sun" can be listened to and read in over 300 versions, in indigenous languages. There is thus a scientific and heritage dimension in this work, insofar as a number of regional or minority languages are in a critical situation.

Keywords: geolinguistics, dialectology, speaking atlas, indigenous languages

Résumé

L'objectif de ce travail est de montrer et de valoriser la diversité linguistique de la France à travers des enregistrements recueillis sur le terrain, une réalisation informatique (qui permet de visualiser les aires dialectales) et un travail de transcription orthographique. Nous décrivons ici un site web (https://atlas.limsi.fr) présentant des cartes interactives de France hexagonale et des Outre-mer, à partir desquelles la fable d'Ésope « La bise et le soleil » peut être écoutée et lue dans plus de 300 versions, en langues locales. Il y a ainsi une dimension à la fois scientifique et patrimoniale à ce travail, dans la mesure où un certain nombre de langues régionales ou minoritaires sont en situation critique.

1. Introduction

Language allows us to communicate but also to reflect our identity. A uniform language where, as in the mythical Tower of Babel, everyone would designate a brick or mortar in the same way, without metaphors, without ambiguity or affectivity, without polysemy, would undoubtedly have a technical utility. The argument that, if we all spoke French or English (or Globish), understanding would be easier, is difficult to counter. It was already in place in French revolutionary Abbé Grégoire (1794), in his report on the need and means of annihilating the "patois", from which we will quote an extract (without translating it).

> Proposerez-vous […] des traductions ? Alors vous multipliez les dépenses […]. Ajoutons que la majeure partie des dialectes vulgaires résistent à la traduction, […] les uns […] sont absolument dénués de termes relatifs à la politique ; les autres sont des jargons lourds et grossiers, sans syntaxe déterminée, parce que la langue est toujours la mesure du génie d'un peuple.

Today, fortunately, we no longer express ourselves in these terms; we are rather saddened by the death of languages, as well as by the disappearance of animal species. We will not push this parallel made by Hagège (2002) with living species too far, because language is above all a social construction. However, a world with no regional particularities would be boring and lack poetry; it would lack flavour like a meal without salt, without pepper. Nowadays, most people are attached to the diversity of languages, which all teach us something about Mankind. As Cyrulnik (2019) says:

> Il faut que le langage soit énigmatique afin de laisser place à l'interprétation. Un langage précis ne serait que désignation, signal de la chose, sans vie émotionnelle, sans vibration, juste une information pour déclencher la réponse.

The question of the linguistic norm quickly arises, as soon as one is interested in the diversity of languages and variation within languages. We can distinguish at least two types of norms: a statistical norm (objective, established by observable and quantifiable facts) and a prescriptive norm (subjective, which indicates a model promoted as "correct"), a cultural construct made up of social choices (Canguilhem, 1966; Gadet, 2003; Rastier, 2007). For many minority languages, in the absence of a prescriptive norm accepted by all, a great deal of variation comes into play. This is understandable, but it is a difficulty we quickly encounter when we go and do field work, for example asking speakers to translate the fable attributed to Aesop "The North Wind and the Sun", a text which has been used for over a century by the International Phonetic Association (IPA) — this story can be accessed in a hundred versions on the IPA website.

In the following, we elaborate on this fable, in order to highlight and promote linguistic diversity: we describe a speaking atlas which takes the form of a website presenting interactive maps of France and its Overseas Territories, where one can click on more than 300 survey points to listen to translations of this text, in regional or minority languages, and read a transcript of what is said. In total, between 2016 and 2018, around 60 languages were collected, half of them in Oceania. These languages do not have unanimously recognised and accepted standards (Caubet et al., 2002). Thus, the transcription solutions proposed vary from one language to another, even among the territorial languages of France, the spellings of which are based on the Latin alphabet and are meant to be close to the pronunciation (at least to some extent). The orthographies adopted are more or less phonetic (reflecting a particular local pronunciation) or diasystemic (emphasising the unity of a set of dialects). Sometimes, the

system is hybrid, for the interests of efficiency, noting what in pronunciation differs from French, while following the orthographic conventions of French. These systems, which have their advantages and disadvantages, are illustrated in the languages addressed here. Fixing the spelling of an indigenous language is an object of research per se, which may not be enough to reverse language shift (Fishman, 1991). However, in this rapidly changing world, in our society where computer- and smartphone-mediated communication holds a preponderant place, we believe that there is hardly any other way for the survival of many minority languages than learning how to write them.

2. Materials

2.1. Languages Collected

Aesop's fable (120 words in French, about 1 minute of speech) was recorded in Basque, Breton, Alemanic Alsatian, Franconian, West Flemish and Romance languages (Boula de Mareüil et al., 2018: see also references therein). It was then translated in the many languages of the French Overseas Territories (Caribbean, Pacific and Indian Oceans): the languages of Guiana (Goury and Migge, 2003; Léglise and Migge, 2007; Biswana, 2016), Creoles (Cayol et al., 1984–1995; Bernabé, 2005; Le Dû and Brun-Trigaud, 2011; Corn, 1990; Ehrhart, 1993), Kanak languages (Rivierre, 1979; Wacalie, 2013; Moyse-Faurie, 2014), Polynesian languages (Lazard and Peltzer, 2000; Vernaudon, 2017) and the languages of Mayotte (Laroussi, 2010). It was also translated in the so-called non-territorial languages of France such as R(r)omani (Courthiade, 2007, 2013) and the French sign language (LSF), with respect to which the French State acknowledges a patrimonial responsibility (Cerquiglini, 1999).

The website (https://atlas.limsi.fr) opens with Hexagonal (i.e. Metropolitan) France, divided in the 25 dialect areas listed below and in Figure 1. Another tab opens a map of the world, which allows navigation from creole to creole and gives access, by clicking inside various rectangles, to additional maps: the American-Caribbean Zone (Antilles and Guiana), the Indian Ocean (Mayotte and Reunion Island), the Pacific Ocean (New Caledonia and Wallis-and-Futuna, on the one hand, French Polynesia on the other). Also, a tab opens a map of the Euro-Mediterranean area, with the non-territorial languages of France. In summary, the languages collected are:

- **Romance languages:** *Oïl* (Picard, Gallo, Norman, Mainiot, Angevin, Poitevin-Saintongeais, Berrichon-Bourbonnais, Champenois, Burgundian, Franc-comtois, Lorrain and Walloon), *Oc* (Gascon, Languedocian, Provençal, North-Occitan and *Croissant* 'Crescent'), Catalan, Corsican and Francoprovençal, with particular signage for Ligurian dialects confined to isolated towns like Bonifacian;
- **Germanic languages:** Alsatian, West Flemish, Franconian (with its Luxembourgish, Mosellan and Rhenish dialects);
- **Breton** (a Celtic language, with its Trégorois, Léonard, Cornouaillais and Vannetais dialects);
- **Basque** (a linguistic isolate, with its Lapurdian, Lower Navarrese and Souletin dialects);
- **French-based creoles**: Guadeloupean, Martinican, Guianese creoles, Reunion creole (from the lowlands and the highlands) and Tayo (the creole of New Caledonia);
- **Nengee languages** (English-based creoles, possibly influenced by Portuguese, of the descendants of Maroons who escaped from slavery in Suriname): Aluku, Ndyuka, Pamaka, Saamaka;
- Hmong (an Asian language brought to the French Guiana) and **Indigenous languages of America**: Kali'na, Wayana (Cariban languages), Arawak, Palikur (Arawakan languages), Teko, Wayãpi (Tupi-Guaraní languages);
- **Mayotte languages**: Shimaore (Bantu), Kibushi (of Malagasy origin);
- **Kanak languages**: Nyelâyu, Jawe, Nêlêmwa, Zuanga, Pwaamei, Paicî, Ajië, 'orôê, Xârâciùù, Drubea, Numèè, Kwényï, Iaai, Drehu, Nengone:
- **Polynesian languages**: Faga Uvea, Wallisian, Futunian, Tahitian (including in its Reo Maupiti variety), Pa'umotu (in its Napuka, Tapuhoe, Parata and Maragai varieties), Rurutu, Ra'ivavae, Rapa, Marquesan (in its 'eo 'enana mei Nuku Hiva, 'eo 'enana mei 'Ua Pou, 'eo 'enata varieties), Mangarevan;
- **non-territorial languages of France**: dialectal Arabic (Moroccan, Algerian, Tunisian, Syrian, Palestinian), Berber (Tashlhiyt, Tamazight and Kabyle), Judeo-Spanish (in its Ḥaketía and Djudyó varieties), Yiddish, Western Armenian, Rromani, LSF (with an audio-visual recording which "doubled" in French by a researcher specialised in LSF, who also wrote an explanatory text about the body gestures which characterise this language).

Options enable the display (or not) of administrative boarders, the legend, the seas and new recordings (as well as their transcripts) outside of France: in Norman (in Jersey), Walloon (in Belgium), Catalan, Aranese Occitan, Aragonese, Basque, Asturian and Galician (in Spain), different Ligurian dialects of Italy, Haitian and Saint-Lucian (French-based) creoles, Bislama (English-based creole from Vanuatu), Fijian, Malagasy, Latin and even Esperanto. A specific checkbox allows the user to zoom on the Crescent, in the centre of France, to display survey points that would otherwise be too close to one another at the scale of the French territory, in and around this area — the limits of which appear fuzzy to highlight the transitional nature of this zone, between *Oïl* and *Oc* languages. Finally, a double orthography has been added for some varieties, in particular Berber (in Tifinagh and Latin alphabets), Western Arminian, Yiddish and Arabic dialects. A page "About" enables the visitor to know more about the ongoing project with some of our publications (Boula de Mareüil et al., 2018, 2019) and to download the data under a Creative Commons license.

2.2. Speakers and Protocol

The speakers we selected, most often elderly people (average age = 60), were from varied socio-professional backgrounds. They were recorded in quiet rooms and asked to sign consents for free distribution. A common protocol was applied, in which the speakers were asked to translate this story into their indigenous language, either directly

with the French text in front of them, or from a written text they preferred to read. The recordings were then intensity equalised, and great care was given to their orthographic transcriptions, which were checked by linguists.

Sometimes the speakers' productions moved away from literal translations to get closer to oral traditions. Some words gave rise to particularly enriching conversations:

- ***bise* 'north wind'**: tra(ns)montane in some Romance dialects, trade winds in some Polynesian dialects, Hebrew loans in some Jewish languages;
- ***voyageur* 'traveller'**: pilgrim in some Romance dialects, wanderer in some Germanic dialects;
- ***manteau* 'coat':** burnous in some Arabic dialects, rain protection in some indigenous languages of America; linen (*jii*) offered to "make the custom" or overcoat (*paleto*, actually a loan word of obscure origin) with which a traditional chief is clothed during his enthronement in some Kanak languages (see the illustration in Table 1).

In all cases, we paid a lot of care to the orthographic transcription of what was said, with in square brackets idioms of beginning (such as 'once upon a time') and idioms of end (of the type 'it is finished') which our Kanak speakers often wanted to add, The different translation strategies testify to the wealth of our linguistic heritage. It seems that this richness is of interest to the general public, in the sense that our site enjoyed a great success in print, broadcast and social media: it has attracted over 600,000 visits since its launch in 2017.

3. Conclusion and Future Work

3.1. Discussion

The aim, through this short outline of our speaking atlas, was to valorise the linguistic diversity of France, relying on a comparable basis which has didactic scope. This is almost urgent, inasmuch as a large number of indigenous languages of France are endangered. With this linguistic atlas, we hope to give prestige to dialects, to give them a positive image, for lack of being able to reverse the decline in their use — transmission among young people is not assured in many cases. It is probably inevitable that most dialects and minority languages of France will be supplanted by a more widely used language like French — which is also mortal. At a time when linguistic diversity and biological diversity are threatened, our profession of faith is that we will devote all our energy to delay the deadline. It is not (only) a matter of folklore tinged with exoticism and essentialism, reifying an idealised past (Bucholtz, 2003). Each language, each dialect provides formal means to express nuances of thought; each language, each dialect refers to a whole imaginary through what words evoke, through the play of sounds. Living with several languages opens up to the Other, it makes it possible to understand difference, it teaches people about the multiplicity of worldviews.

In the future, we plan to expand our linguistic mapping activity, combining field surveys and crowdsourcing: we still aim at enriching the speaking atlas of the languages and dialects of France, which on the other hand we recently extended to Italy, Belgium and Spain. The objective is to develop its educational aspect in two directions: starting from the existing materials, to which glosses as in Table 1 should be added, and from new surveys, two case studies will be considered (Romance languages and Polynesian languages) to immediately grasp variation through audio paths and the visualisation of isoglosses. Isolated words and/or linguistic notions will be presented, as in traditional dialectological atlases (Gilliéron and Edmont, 1902–1910; Charpentier and François, 2015). Visually and acoustically (the data are currently being transcribed phonetically by using forced alignment), different types and groupings of dialects will be illustrated.

3.2. Linguistic Mapping, Crowdsourcing and Dialectometry

Innovative visualisation methods will be developed, centered on the Gallo-Romance area on the one hand, on French Polynesia on the other hand — Romance and Polynesian languages presenting a certain homogeneity within each family. Linguistic variables will be selected: lexical items such as the words/concepts *wind*, *sun* and *cloak*, for example, taken from the recordings of the Aesop fable already mapped in more than 300 translations in the current version of our speaking atlas, to which words/concepts will be added from the *Atlas Linguistique de la France* (Gilliéron and Edmont, 1902–1910) and the Swadesh list collected during new surveys (Swadesh et al., 1971). The user will be able to follow phonetic and lexical changes by simply moving the mouse, possibly following courses which will be proposed. Isoglosses will also be generated more or less automatically, using clustering techniques, the results of which will be projected on maps, with different colour codes associated with different types and groupings of dialects. For these different aspects, we will also offer remote recording/transcription methods to enrich the database. In the longer term, solutions will be considered to go further (with one or more hundreds of words/concepts) and make the site sustainable.

The web will be used in order to display the fruits of research, as well as to collect new information, using a crowdsourcing methodology. This type of methodology will be particularly appreciable for French Polynesia, to avoid travelling in archipelagos very far from each other. Completing remote survey points, though, requires caution: we will take care of the linguistic content of the recordings and their transcripts. The breadth of the territory covered by this project and the diversity of the languages represented (Romance or even neo-Romance languages such as French-based creoles and Oceanian languages) opens up other avenues of research, which will require new collaborations.

In addition to the functionalities of the https://atlas.limsi.fr site, an attractive interface will also guide the journey through space, in Hexagonal France and French Polynesia — where extremely interesting lexical taboo phenomena (*pi'i*) exist. In order to go beyond the Aesop fable "The North Wind and the Sun" and to integrate a list of isolated words, for fauna and flora, for example — reviving with

the traditional practice of linguistic atlases —, we wish to design and/or improve an existing recording tool, LIG-Aikuma (Adda et al., 2016), to help field linguist in their investigations. Such an application will make it possible to capitalise on the network of dialect (or minority language) speakers we have, to ask them to record themselves. The program will need to be user-friendly, as speakers are often rather old. And this smartphone-based approach does not prevent the need for further field surveys, because human relationships are essential and because it may be important to be with the informants to start the process. The task of the field linguist equipped this way will be facilitated, and the processing of the data collected will be faster, to apply various dialectometric techniques.

4. Acknowledgements

This work was largely financed within the framework of the "Langues et Numérique 2016 & 2017" calls for proposals of the Délégation Générale à la Langue Française et aux Langues de France. We are grateful to the Académie des Langues Kanak and to all those who have agreed to give us their time and lend their voices to this achievement.

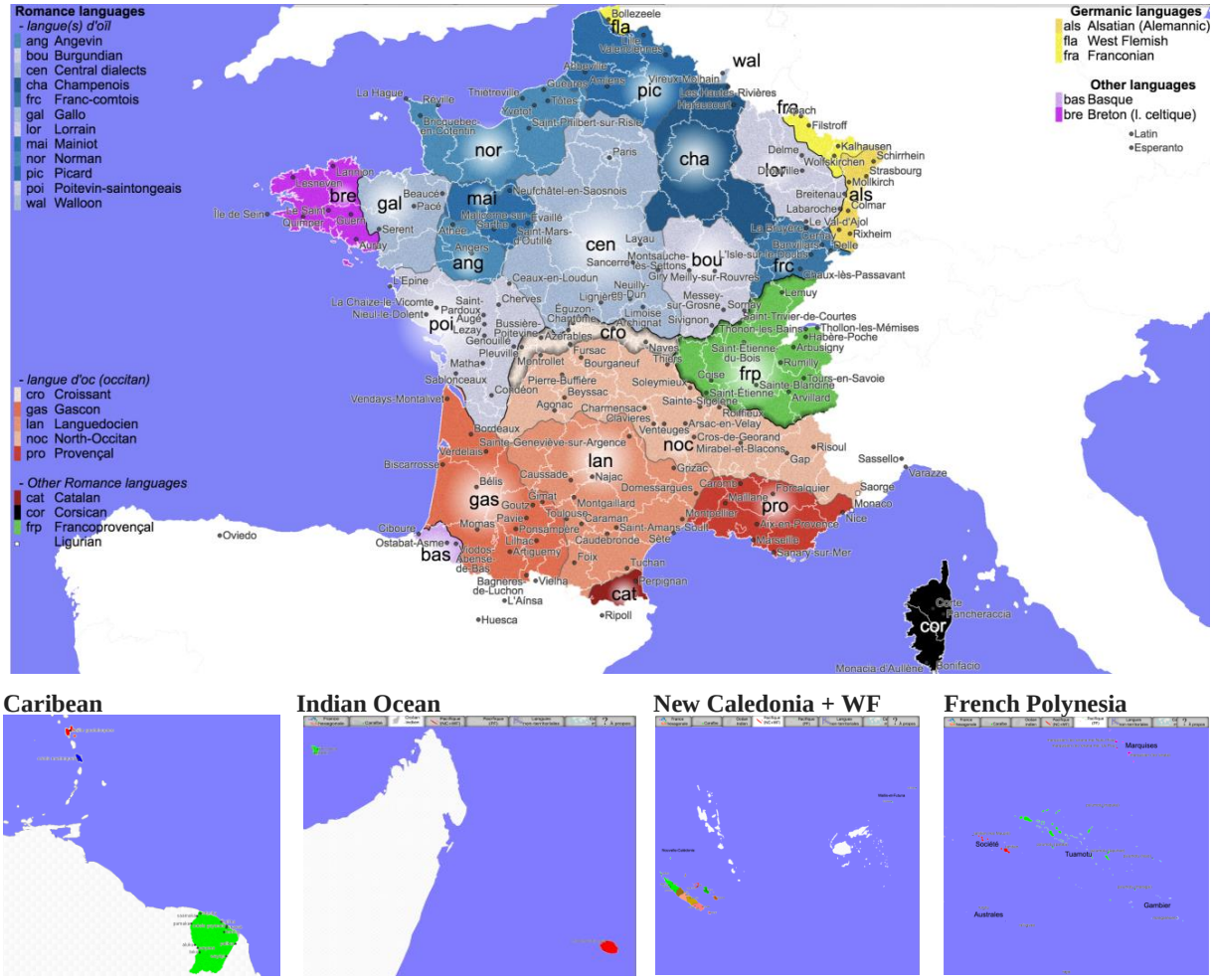


Figure 1: Maps of Hexagonal France and Overseas Territories (WF = Wallis-and-Futuna).

Naa	mwa	tûâ	taa	vinyié	nâ	nyî	vê mé	iitèè	naa	i pwa	ngê	taa	jii.
They two	TAM	see	a	man	and	he	come	meet	they two	covered	with	a	cloth

Nââ	ngê	tôâ	taa	xèxùù	na	yi	vê	bwè	nââ	iévititéé	mô	taa	paleto.
They two	TAM	see	a	man	who	he	come	towards	they two	wrapped	with	a	coat.

Table 1: Excerpt from the Aesop fable with literal translations in two Kanak languages (Numèè and Kwényï) spoken in the Djubéa-Kaponé customary area of New Caledonia. TAM = tense-aspect-mode particle.

5. Bibliographical References

Adda, G., Stüker, S., Adda-Decker, M., Ambouroue, O., Besacier, L., Blachon, D., Bonneau-Maynard, H., Godard, P., Hamlaoui, H., Idiatov, D., Kouarata, G.-N., Lamel, L., Makasso, E.-M., Rialland, A., Van de Velde, M., Yvon, F., Zerbian, S. (2016). Breaking the Unwritten Language Barrier: The BULB Project, *Procedia Computer Science*, 81, 8–14.

Bernabé, J. (2005). Guadeloupe et Martinique : Un survol sociolinguistique. *Langues et cité*, 5:6–7.

Biswana, H. (2016). Luta cultural e por direitos: reflexões dos Arowaka sobre a cultura, *Boletim informativo*, 2:22–26.

Boula de Mareüil, P., Vernier, F., Rilliard, A. (2018). A Speaking Atlas of the Regional Languages of France, *Proceedings of the 11th International Conference on Language Resources and Evaluation*, Miyazaki, 4133–4138.

Boula de Mareüil, P., Adda, G., Lamel, L., Rilliard, A., Vernier, F. (2019). A speaking atlas of minority languages of France: collection and analyses of dialectal data. *Proceedings of the 19th International Congress of Phonetic Sciences*, Melbourne, 1709–1713.

Bucholtz, M. (2003). Sociolinguistic nostalgia and authentication of identity. *Journal of Sociolinguistics*, 7(3):398–416.

Courthiade, M. (2007). Jeu dialectes-langue. *Langues et cite*, 9:6–7.

Courthiade, M. (2013). *A succint history of the Rromani language*. INALCO, Paris.

Canguilhem, G. (1966). *Le normal et le pathologique*. Presses Universitaires de France Paris.

Caubet, D., Chaker, S., Sibille, J. (2001). *Codification des langues de France.* L'Harmattan, Paris.

Cayol, M., Chaudenson, R., Barat, C. (1984–1995). *Atlas linguistique et ethnographique de la Réunion*. Éditions du CNRS, Paris.

Cerquiglini, B. (1999). Rapport au Ministre de l'Éducation Nationale, de la Recherche et de la Technologie, et à la Ministre de la Culture et de la Communication <http://www.ladocumentationfrancaise.fr/var/storage/rapports-publics/994000719.pdf>.

Charpentier, J.-M. and François, A. (2015). *Atlas linguistique de la Polynésie française*. Mouton de Gruyter, Berlin.

Corn, C. (1990). Tayo pronouns: a sketch of the pronominal system of a French-lexicon Creole language of the South Pacific. *Te Reo*, 33:3–24.

Cyrulnik, B. (2019). *La nuit, j'écrirai des soleils*, Éditions Odile Jacob, Paris.

Ehrhart. S. (1993). *Le créole français de St-Louis (le tayo) en Nouvelle-Calédonie*. H. Buske Verlag, Humburg.

Fishman J. A. (1991). *Reversing language shift: Theoretical and empirical foundations of assistance to threatened languages*. Multilingual Matters, Clevedon.

Gadet, F. (2003). *La variation sociale en français*. Ophrys, Paris.

Gilliéron, J. and Edmont, E. (1902–1910). *Atlas linguistique de la France*. Champion, Paris.

Goury, L. and Migge, B. (2003). *Grammaire du nengee. Introduction aux langues aluku, ndyuka et pamaka*. IRD Éditions, Paris.

Grégoire, Abbé H. (1794). *Rapport sur la nécessité et les moyens d'anéantir les patois et d'universaliser l'usage de la langue française*. Convention nationale, Paris.

Hagège, C. (2002). *Halte à la mort des langues*. Odile Jacob, Paris.

Laroussi, F. (2010). *Langues, identités et insularité : Regards sur Mayotte*. PURH, Rouen/Le Havre.

Lazard, G. and Peltzer, L. (2000). *Structure de la langue tahitienne*. Peeters, Paris.

Le Dû, J. and Brun-Trigaud, G. (2011). *Atlas linguistique des petites Antilles.* Éditions du CNRS, Paris.

Léglise, I. and Migge, B. (2007). *Pratiques et représentations linguistiques en Guyane.* IRD Éditions, Paris.

Moyse-Faurie, C. (2014). L Nouvelle-Calédonie et le statut des langues kanak : quelques repères historiques. *Langues et société*, 26:9–11.

Rastier, F. (2007). Conditions d'une linguistique des normes. In G. Steuckard, S. Sioufi (editors), *Les linguistes et la norme — Aspects normatifs du discours linguistique*, Peter Lang, Bern (pages 3–20).

Rivierre, J.-C. (1979). Langues de l'extrême-sud et plus particulièrement le nââ kwênyìì, langue de l'Île des Pins. In A.-G. Haudricourt, J.-C. Rivierre, F. Rivierre, C. Moyse-Faurie, J. de la Fontinelle (editors), *Les langues mélanésiennes de Nouvelle-Calédonie*. Nouméa, DEC (pages 72–79).

Swadesh, M., Sherzer, J., Hymes, D. H. (1971). *The Origin and Diversification of Language*. Aldine, Chicago.

Vernaudon, J. (2017). L'origine des langues polynésiennes. *Langues et cité*, 28: 2–3.

Wacalie, F. S. (2013). *Description morpho-syntaxique du nââ numèè (langue de l'extrême-Sud, Nouvelle-Calédonie)*. Phd thesis, INALCO, Paris.

Proceedings, pages 160–163
Paris, UNESCO Headquarters, 5-6 December, 2019.

ELLORA: Enabling Low Resource Languages with Technology

Kalika Bali, Monojit Choudhury, Sunayana Sitaram, Vivek Seshadri
Microsoft Research Labs
Bangalore, India
{kalikab, monojitc, susitara, visesha}@microsoft.com

Abstract

Language technology has had a huge impact on the way language communities communicate and access information. However, this revolution has bypassed over 5000 languages around the world that have no resources to develop technology in their languages. ELLORA, with its mission to empower every person and every organization using underserved languages to achieve more, is a program for enabling low resource languages through language technology. In this paper we describe how through innovative methodologies and techniques that allow systems to be built in resource constrained settings, ELLORA seeks to positively impact the underserved language communities around the globe..

Keywords: Low resource languages, language technology, language data, endangered and minority languages

Résumé

भाषा प्रौद्योगिकी के माध्यम से डिजिटल क्रांति की अगुवाई में भाषा समुदायों को जानकारी पहुँचाने के तरीके पर भारी प्रभाव पड़ा है। जबकि कुछ प्रमुख संसाधन संपन्न भाषाओं के उपयोगकर्ता हर दिन ऐसी तकनीक का लाभ उठा रहे हैं, इस क्रांति से दुनिया भर की अधिकांश भाषाएँ उपेक्षित हैं । प्रौद्योगिकी विकास की डेटा-भूखी दुनिया में, 5000 से अधिक भाषाओं को अपनी भाषाओं में प्रौद्योगिकी विकसित करने के लिए कोई संसाधन नहीं हैं। एलोरा (ELLORA) का मिशन प्रत्येक व्यक्ति और प्रत्येक संगठन को अपनी भाषा में आगे बढ़ने के लिए प्रोत्साहित करना है । यह प्राकृतिक भाषा प्रसंस्करण, वार्तालाप और भाषण प्रौद्योगिकी के माध्यम से कम संसाधन भाषाओं को सक्षम करने के लिए एक कार्यक्रम है। इस शोध पत्र में हम बताते हैं कि कैसे नवीन पद्धतियों और तकनीकों के माध्यम से ELLORA दुनिया भर के भाषा समुदायों को सकारात्मक रूप से प्रभावित करना चाहता है।

1. Introduction

Technology pervades all aspects of society and continues to change the way people access and share information, learn and educate, as well as provide and access services. Language is the main channel through which such transformational technology can be integrated into the socio-economic processes of a community.

However, this benefit is still limited to a subset of the world's language communities and large populations worldwide are bereft of access to technology in their own languages. Most languages in the world lack the linguistic resources to build large data-driven (e.g., Deep Neural Net) models. To be able to truly support speech and language systems that can enable everyone on the planet, methodologies and techniques to build systems in resource constrained settings are essential.

ELLORA-a program for enabling low resource languages through Natural Language Processing, Conversation and Speech technology was established at Microsoft with the view to address the needs and aspirations of language-users currently unable to access such technologies. ELLORA aims to impact underserved communities through enabling language technology by creating economic opportunities, building technological skills, enhancing education and preserving local language and cultures for future generations. The approach taken is to start with the low resourced languages of India and then scale it to the low resource languages in the rest of the world.

In recent times, there have been a number of breakthroughs in the field of NLP, which is primarily due to the use of deep neural networks and availability of large amount of language resources as well as computational power. Languages resources include both linguistic datasets that are used for training language processing systems, and basic language processing tools such as stemmers, morphological analyzers, parsers etc., which enable other language processing technologies.

In September 2017, Microsoft announced a speech recognition system that could achieve better-than-human performance in speech transcription (Xiong et al., 2017), which used 200M transcribed words from conversational speech. In 2018, human-parity was achieved for English-Chinese Machine translation, again training on 18M bilingual sentence pairs (Hassan et al., 2018). These achievements hold a lot of promises, particularly in making information and technology accessible to the speakers of a language. Unfortunately, these technologies work only when there is large amount of training data. For most languages in the world, there is hardly any resources available.

In a 2008 study of the Linguistic Data Consortium (LDC) portal, Choudhury (2008) shows that resource distribution across languages follow power-law, with only four languages – Arabic, Chinese, English and Spanish – having a large amount of resources and only a handful having some, leaving 90% of the world languages in the long tail of this distribution with hardly any resources (Fig 1) to train any useful NLP system. Choudhury (2008) predicts that since availability of language resources propels creation of technology and more resources, and attracts more

researchers and technologists toward that language, this digital divide between languages will widen further with time, unless intervened aggressively and strategically. Indeed, today, after a decade, while technology has grown more and more data intensive, LDC catalogue does not list any resource for Javanese and Kannada, which are the 12th and 32nd most spoken languages with 98M and 43.7M speakers respectively.

A study by Caribou Digital Analysis (Will et al., 2019) summarizes the economic aspect of this divide as follows: "Analysis shows there is a clear income gap in access. Google Translate is available in the languages spoken by just 54% of those living on less than $1.90 per day. The picture is even more stark for Natural Language Understanding frameworks such as Dialogflow, which supports languages spoken by only 3% of those living on less than $1.90 per day."

Based on the availability of digital resources and access to technology, the languages of the world can be broadly classified into four groups, as illustrated by the pyramid in Fig 1. Only the top of the pyramid, which has 10-15 economically important languages, are positively affected by the breakthroughs in AI and NLP. Research and development of technology for low resource languages have always been in the fringes. The recent advances in language technology are mostly beneficial to languages in the second and third tier, which together cover another 100 languages. The remaining 5000+ languages, i.e., the bottom of the pyramid, have no resources, and consequently the speakers do not have access to technology in their native languages.

2. Enabling Data and Technologies

Language technology enables access to information as well as broader technology, such as through local language smart phone interfaces. This can be viewed as a multi-layer approach to enabling a language, where the more complex AI technologies build upon fundamental and simpler layers. Input/Output Support form the first layer of language technology, that include basic font and keyboard support, and more advanced features like text prediction, and spelling and grammar correction. Speech input/output mechanisms include speech-to-text and text-to-speech systems and are more relevant for languages without script and where a large fraction of low-literate users. Local language UI is the next layer of support, which ranges from availability of particular OS or apps in a language to generic technological support for building UI in a language. Information access is enabled by text and voice based search technologies. However, languages that have little content on the Web will particularly benefit from machine translation techniques that can be used to translate content from other languages. Finally, digital assistants and conversational interfaces are useful for ease of interfacing with the devices, various other technologies and are enablers of businesses in that language.

Building some of these technologies require linguistic expertise, while others are data intensive. Some technologies, such as speech-to-text, require transcribed speech data, which is labor intensive and requires native speakers to label/transcribe data, whereas other technologies such as text prediction can be built by training on any corpus of the language and requires no further human labeling or processing. This makes some of the technologies more expensive, time intensive and challenging to build for low resource languages. Table 1 summarizes the data/expertise requirement for various technologies and their availability.

As can be seen from the table, technologies that require high to moderate amounts of labeled data are typically unavailable for No and Low resourced languages. Languages in the bottom of the pyramid also lack technologies which require only unlabeled data. This is not surprising as creation of labeled data is expensive and there may not be any financial incentive to invest in these languages. While Low-resourced languages have the basic enabling technologies, they are deprived of high-quality advanced tools, especially those that require labeled data

3. ELLORA Data Initiatives

As discussed in the above sections, building language technologies requires significant data resources in the desired language. As a commercial initiative, there has been little interest in developing these resources for indigenous and minority languages, as well as the languages of the Global South. Data thus, remains the single biggest bottleneck as language technology models become more and more data hungry. Data remains at the heart of several activities undertaken through Ellora, ranging from innovative data collection initiatives to providing seed data for nurturing research.

Microsoft India released speech training and test data for Telugu, Tamil and Gujarati in Sept 2018. This is the largest publicly available Indian language speech dataset aimed at helping researchers and academia build Indian language speech recognition for all applications where speech is used. This data was used as the basis of the Low Resource Speech Recognition Challenge, held as a part of Interspeech 2018. Participants were able to create high quality speech recognition models using the Microsoft Indian language speech corpus, thus validating the efficacy of the corpus (Srivastava et al, 2018)

As data collection remains an expensive exercise, an effort to collect high quality data at low cost is one of the goals for ELLORA. Karya (Chopra et al, 2019) a crowdsourcing platform to provide digital work to low-income workers, presents an ideal vehicle for this. The motivation for Karya lies in the fact that roughly half the Indian population lives under $5 a day. By identifying training needs and imparting digital literacy to low-income population, Karya enables microtasking in an efficient manner on a mobile platform. Thus, it significantly increases the current daily wage of the workers while at the same time reduces the cost of task completion which may allow data collection at scale. Currently, an experimental project in collaboration with Indian Institute of Technology, Bombay, to determine the feasibility of using Karya successfully for speech data collection is underway in rural and semi-urban parts of the Indian state of Maharashtra. While, the final results are awaited, the initial feedback seems encouraging.

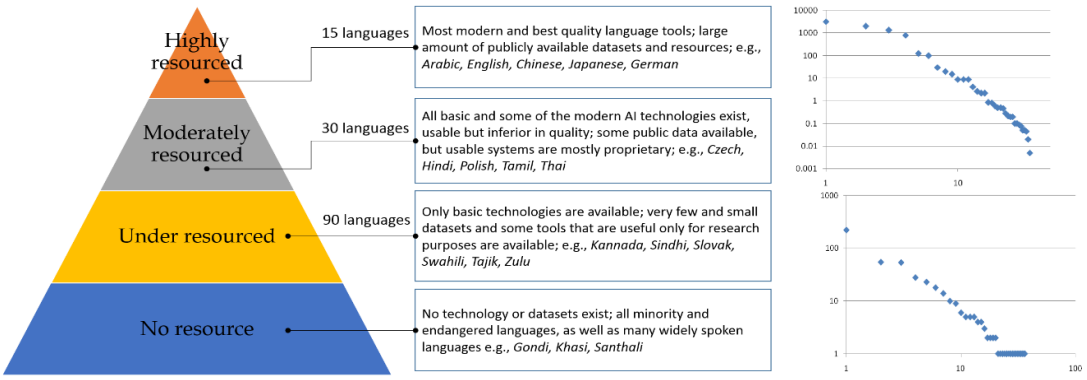


Figure 1: Classification of languages according to the availability of language technology, tools and resources (left) based on the power-law distribution of the resources across the languages of the world (right).

4. Enabling Minority Languages

Some of the most disadvantaged socio-economic groups in the world are also the ones with the least access to language technology. ELLORA's impact can only be measured if the communities using languages with no technological support can be provided access through technology in their own language. To understand the viability of such a social impact, ELLORA is working with CGNet Swara on building and deploying Gondi language technology.

Gondi is a South-Central Dravidian language and is in the 'vulnerable' category on UNESCO's Atlas of the World's Languages in Danger (Mosley, 2010). Spoken by nearly 3 million people (Indian Census, 2011) in the Indian states of Chhattisgarh, Andhra, Odisha, Maharashtra and Karnataka, it is heavily influenced by the dominant state language. CGNet Swara provides a citizen journalism portal for the tribal regions of Chhattisgarh and home to the Gondi language community, by making local stories accessible through mobile phones. As there is absolutely no language technology support for Gondi, most of the content is created, moderated and edited manually. Targeted language technology applications can increase the scale, and hence the access to information for a community that is marginalized and lives in areas of civil unrest.

A meeting was held in April 2018 to understand the potential impact of Language Technology for Gondi on the community and brainstorm on transformational technology applications. The discussions involved stakeholders and experts from CGNet Swara, academic institutes like IIIT Raipur, IIT-KGP, Jadavpur University etc, Microsoft and other non-profit organizations like Pratham Books. Subsequently, a workshop was organized at IIIT Naya Raipur in collaboration with Pratham Books , CGNetSwara and Microsoft Research. The Gondi speakers who participated in the workshop translated approximately 200 books on Storyweaver from Hindi to Gondi. Not only was this the first step towards creating parallel data for Gondi-Hindi that can be of use in building Machine Translation systems for Gondi, it made available children's books for the first time in the language

Adivasi Radio, a Mobile News App for Gondi has also been developed through this collaboration. The first version released uses Text-to-Speech synthesis in Gondi to read out news and articles available on the CGNetSwara site on the users' phones. Future development envisages the incorporation of a Machine Translation system that allows news articles and other content in Hindi to be translated and read out in Gondi. This would have a major impact on the community by providing access to news in local language, while also producing content in Gondi.

Technology	**Availability of technology for the resource status of a language**				**Data/Expertise Requirement**		
	High	**Mode-rate**	**Low**	**No**	**Linguistic Expertise**	**Unlabel-ed Data**	**Label-ed Data**
Input/Output Support							
Font & keyboard	☆☆☆	☆☆☆	☆☆☆	☆☆	☆☆☆		
Speech-to-text	☆☆☆	☆☆			☆	☆☆	☆☆☆
Text-to-speech	☆☆☆	☆☆	☆		☆☆☆		☆☆
Text prediction	☆☆☆	☆☆☆	☆☆			☆☆☆	
Spell checker	☆☆☆	☆☆☆	☆☆		☆☆☆	☆☆	
Grammar checker	☆☆☆	☆☆			☆☆	☆☆☆	☆☆
Local language UI							
	☆☆☆	☆☆☆	☆		☆☆☆		
Information access							
Text search	☆☆☆	☆☆	☆		☆	☆☆☆	☆☆
Voice to Text search	☆☆☆	☆				☆	☆☆☆
Voice to speech search	☆☆	☆			☆	☆☆☆	☆☆☆
Machine translation	☆☆	☆	☆		☆☆		☆☆☆
Conversational systems							
	☆☆	☆			☆☆☆	☆☆☆	☆☆☆

Table : Enabling language technologies, their availability and quality

5. Discover, Design, Develop and Deploy

While it is clear that a large number of languages in the world require intensive investment in resource creation for technology enablement, it seems highly unlikely that such an investment can be delivered readily and easily in a short span of time. It is thus imperative that the investment is done in a manner that ensures maximum benefit for a community through language technology. To enable language technology to deliver a positive social impact on the low-resource language communities around the world, we propose the use of a modified version of the 4-D design thinking methodology of Discover, Design, Develop and Deploy.

Discover the problem that language technology can address for a particular language community. This observation led approach can help target resources where they are most needed.

Design for the users and their language. Understand the diversity in linguistic properties of languages and their usage. Avoid a majority language (usually English) led approach where all effort is spent on adapting an existing technology for a dominant language. The cost of ignoring pertinent language features in such a process often results in a less than optimal technology development.

Develop rapidly and Deploy frequently. An iterative process can ensure early failures lead to success.

Such a user-centric approach will not only deliver a more functional language technology but also ensure a more equitable and beneficial distribution of resources.

This approach, at the heart of all ELLORA activities, is well-illustrated in a project on Mundari language. This collaboration between ELLORA and IIT-Kharagpur has a team of scholars led by Prof Dripta Piplai and Prof Manjira Sinha have been spending time with the Mundari-speakers in the Jhargram villages in the eastern state of West Bengal observing and recording the nuances of the language and its use (Mitra, 2019) This information will feed into a mobile based app for teaching, learning and communication in Mundari. According to Dr Piplai, "The close interactions have thrown up interesting facts. One of them is that the younger generation does not use it in its pure form. They communicate in a mix of Bengali, Mundari and Oriya"

The digital revolution spear-headed by language technology has had a huge impact on the way language communities communicate and access information. However, this revolution seems to have bypassed the 5000+ languages at the bottom of the pyramid with zero resources available to them. ELLORA aims to bridge this gap by supporting language technology systems and applications to enable everyone on the planet. Through innovative methodologies and techniques that allow systems to be built in resource constrained settings, ELLORA seeks to positively impact the underserved language communities around the globe.[1].

6. Acknowledgements

We would like to acknowledge Manu Chopra, Shubhranshu Choudhary, Sandipan Dandapat, Rupesh Mehta, Niranjan Nayak, Dripta Piplai, Manjira Sinha, Brij Mohan Lal Srivastava, IIT Bombay, IIIT Naya Raipur, Pratham Books, Gondi workshop participants and the people of Amale village for their contribution to ELLORA.

7. Bibliographical References

Awadalla, H. et al. (2018). Achieving human parity on automatic Chinese to English News Translation. arXiv:1803.05567

Census (20111). Primary Census Abstracts, Registrar General of India, Ministry of Home Affairs, Govt of India

Chopra, M. et al (2019). Exploring crowdsourced work in low-resource setting. ACM CHI Conference on Human Factors in Computing Systems.

Choudhury, M, (2008). Breaking the Zipfian barrier of NLP. Invited Talk. . In the Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages.

Gericke, K and Blessing, L (2011). Comparisons of design methodologies and process models across domains: a literature review. In DS 68-1: Proceedings of the 18th International Conference on Engineering Design (ICED 11), Impacting Society through Engineering Design, Vol. 1: Design Processes, Lyngby/Copenhagen, Denmark.

Mager, M. et al (2018). Challenges of language technologies for the indigenous languages of the Americas. Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, August 2018.

Mitra, D. (2019). IIT set to launch app in Mundari to keep indigenous language relevant. Times of India, February 2019.

Mosley, C. (ed.) (2010). Atlas of the World's Languages in Danger, 3rd edn. Paris, UNESCO Publishing.

Srivastava, B.M.L. et al (2018). Interspeech 2018 Low Resource Automatic Speech Recognition Challenge for Indian Languages. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages 29-31 August 2018, Gurugram, India

Willis, A, Barrie, G, and Kendall, J (2019). Conversational interfaces and the long tail of languages in developing countries. https://dfslab.net/wp- content/uploads/2019/01/NLP-Language-Divide-Report-.pdf

Xiong, W. et al. (2017). Toward Human Parity in Conversational Speech Recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing | September 2017, Vol 25: pp. 2410-2423

Proceedings, pages 164–168
Paris, UNESCO Headquarters, 5-6 December, 2019.

Dependency Parsing Based On Uzbek Corpus

Nilufar Abdurakhmonova
PhD, associate professor of Information technology department at Tashkent State University of the Uzbek language and literature
anilufar@navoiy-uni.uz

Abstract

Syntactic parsing is crucial stage among existing different types of parsing methods in the field of NLP. Syntactic parsing assists to identify the type sentence and word combinations that represented grammatical relations of the words. However, there are various grammatical features of the languages, almost all languages follow common linguistic rules. The Uzbek language belongs to agglutinative language family based on free constituent order language in syntax. Our investigations show that morphological aspect of word forms plays an essential role to identify and compose syntactic relations for the Uzbek language. Given morphological and lexical information can solve the some problems which connecting with syntactic parsing as well. Our article represents some main point of views the stages of parsing on CoNLLU format based on Uzbek corpus analysis.

Tabbiy tilni qayta ishlashda turli tahlil qilish metodlari orasida sintaktik analiz qilish muhim sanaladi. Sintaktik analiz tilning grammatik munosabatlari aks etgan so‘z birikmalari va gap turlarini aniqlashga hizmat qiladi. Tillarning turli grammatik xususiyatlari bo‘lishiga qaramay, barcha tillar deyarli bir-biriga yaqin umumiy lingvistik qoidalariga bo‘ysunadi. O‘zbek tili agglyutinativ tillar oilasiga mansub bo‘lib, uning sintaksisi ancha erkin komponentlardan iborat. Bizning tadqiqotimizda sintaktik tahlil uchun so‘zshakllarni morfologik jihatdan sintaktik munosabatlarni tuzish va turlarini aniqlashda muhim ekanligi o‘z isbotini topgan. Morfologik va leksik ma’lumotlarning berilishi sintaktik tahlildagi lingvistik muammolarni aniqlashga ham yordam beradi. Maqolamizda o‘zbek tili korpusiga asoslangan CoNLLU formatida ifodalangan sintaktik tahlil bosqichlari tahlil qilingan.

Key words: CoNLLU, corpus, syntactic parsing, the Uzbek language

INTRODUCTION

The Uzbek language has rich paper versions of lexical resources. Currently gathering and selecting different types of Uzbek texts as a corpus implemented by Computational linguistics lab at Tashkent State university of Uzbek language and literature. One of general conception of composing computational models of corpus providing the texts is morphological analysis and syntactic parsing. Nevertheless, our corpus is not open available platform for users due to testing still the results of our project.

One is crucial issue for construction of corpus is to create the model that is ready for analyze the text morphologically and syntactically. Computational point of view grammar is more important for corpus driven language analysis. Parsing is a fundamental process in any natural language processing pipeline, since obtaining the syntactic structure of sentences provides us with information that can be used to extract meaning from them: constituents correspond to units of meaning, and dependency relations describe the ways in which they interact, such as who performed the action described in a sentence or which object is receiving the action (Carlos Gymez-Rodrigue, 2010)

As of early September 2018, there are 132 treebanks for 74 languages publicly available at http:// universaldependencies.org/,1 with 15 upcoming treebanks for a further 13 languages. New UD treebanks are often the result of converting corpora adhering to other annotation schemes—not only dependency-based, but also constituency-based (Adam Przepiórkowski , 2016).

THE NATURE OF UZBEK GRAMMAR

Grammar consist of two parts: morphology and syntax. Both of them are important layer of linguistics for NLP. The Uzbek language morphemes derived a number of combinations of word forms by concatenation root and affixes in most cases. Morphotactics of the language not every time follows the rules owing to some exceptions though there are an exact order of word combination. Usual order of the words:

Root+DerAff+Pl+Poss+Case+Particle (kutub-xona-lar-i-ga-mi);

Root+ DerAff+Voice+Neg+Tense+Particle (oq-la-t-tir -ma-di-mi)

The general structure of the Uzbek language of the sentence follows SOV order. It is free constituent order because of all parts of speech nearly depends on the verb, therefore, there is not difference the meaning though changing places of parts of speech. Due to complex structure of the sentences with grammatical morphemes as a sequence of inflectional elements of parts of speech. We can see one example for this:

Men muzeyga sen bilan ertaga boraman.

S^0=[a+b+c+d+f]

Men sen bilan ertaga muzeyga boraman.

S^1=[[a+c+d+b+f]

Men ertaga muzeyga sen bilan boraman.S^2=[a+d+b+c+f]

Even some inflectional groups give opportunity to identify the function of words by morphological markers. [Noun+Case] model as syntactic marker helps to identify the function of the words in the text, but not every time.

Noun+Gen=> Attributive (bolaning – child’s)

Noun+Acc=>Object (bolaga-to the child)

Conducting our research dedicated more on syntactic annotation in order to create the model of corpus analysis.

Our aim is to build Treebank as an example universal dependency structures so that to use them for NLP. The Uzbek language is more specialized morphological features sequences by order adding morphemes. Hence, representation morphological features is crucial for syntactic parsing as well.

Subsequently, we should clarify the forms and POS in the Uzbek language. The grammatical structure of words in Uzbek can be in following forms:

1) morphemes (affix and stem) – *chorvachilik*
2) morphological variations which composed different functions of parts of speech –*kitob + lar + im +ga (to my books)*
3) word combinations as syntactic level – *tug'ilgan kun uchun sotib olmoq*
4) compound words – *mashq qilmoq, sotib olmoq*
5) phrasal units represented as unique components - *yuragi dov bermaslik*

For accuracy of our parser needs morphological analysis. In Uzbek, being ambiguous word structures confuses the type of the grammatical features as example of verb:

[V]+ib+V=>Sotib olmoq-**compound verb**
[V]+ib +V=>Gulib gapirmoq-**adverbial clause**
[V]+ib +V=>Kulib yubormoq-**collacation**

Giving lexicon and rules for each lexical unit allow us to establish their lexical features and combine above pointed types of word forms though they are alike formally. Authors pointed out Word-based Model and IG-based model for choosing parsing units according to grammatical features (Gülşen Cebiroğlu Eryiğit; Kemal Oflazer; Joakim Nivre, 2013).

Hence, we use FST technology to analyze at the first morphological analyzing, thanks to the Helsinki finite state technology it builds the amount of combination of morphemes (Abdurakhmonova N.; Tuliyev U., 2018).

Furthermore, not only morphological categories but also syntactic relations between the words are important to classify the set of sentences in Treebank for Uzbek. Hence, we use tag set to identify each sentence type and word combination through morphological word forms.

WC	Word combination
COLC	Collocation
FP\FCOLC	Free phrase\ Free collocation
NP	Noun Phrase
NA	Noun Adjoinment
NG	Noun Government
NCS	Noun Collateral subordination
VP	Verb Phrase
VA	Verb Adjoinment
VG	Verb Government
AGRM	Agreement
SLP	Singular personal pronouns
PPL	Plural personal pronouns
ICN\CPCT	Interconnectedness\Complicity
S	Simple sentence
Sub	Subject
Obj	Object
Attr	Attributive
Mod	Modifier
Pre	Predicate

Probability of syntactic structures by pure grammatical approach is more complex than statistical approach by corpus. Consequently, it is necessary to be exist the corpus in order to construct dependency tagging.

METHODOLOGY OF PARSING

Universal Dependencies (UD) is a framework for consistent annotation of grammar (parts of speech, morphological features, and syntactic dependencies) across different human languages. Our general workflow of parsing stages represented the following pic. 1.

The corpus consists of hand built selection of Uzbek fiction annotation with metadata respectively by genres. Here grammatical categories are crucial to give representativeness of features of parts of speech.

A special program intersecting composition was developed in order to facilitate the combining of the lexicon transducer and the two-level rule transducers (TWOLC-two-level compiler) and to avoid excessively large intermediate results (Alexandr Rosen, 19).

In order to morphological analysis there are three components of the Uzbek language: alphabet (Latin and Cyrillic), grammatical rules and Lexicon. In Uzbek the following morphotactics of words as example of Noun:

```
LEXICON NumC
  +SG:    Poss1;
  +PL:lar  Poss2;
LEXICON NumV
  +SG:    Poss2 ;
  +PL:lar  Poss2;
LEXICON Poss1
 +PP1+PSG:m     Case ;
 +PP2+PSG:ng    Case ;
 +PP3+PSG:si    Case ;
 +PP1+PPL:miz   Case ;
 +PP2+PPL:ngiz  Case ;
 +PP3+PPL:i     Case ;
 0:0            Case ;
LEXICON CyrPrePrefinal1
 0:0            Final ;
 +PART:ми       Final ;
 +PART:ку       Final ;
 +PART:-^Ya     Final ;
 +PART:-да      Final ;
 +PART:-чи      Final ;
```

The algorithm of analysis represented Fig.1.

We apply Turkish model to analyze the texts for CoNLLU format, hence there have been the sharp distinction between Turkish and Uzbek structures, but thank to by human correction, grammatical features tagging improved according to

newpar

sent_id = 266
text = Shoir yigitga dil-dildan achinarkan , uni ilk bor uchratgan paytini esladi .
1 Shoir Shoir NOUN Noun Case=Nom|Number=Sing|Person=3 2 nmod _ SpacesAfter=\r\n
2 yigitga yigitga NOUN Noun Case=Nom|Number=Sing|Person=3 3 nmod _ SpacesAfter=\r\n
3 dil dil NOUN Noun Case=Nom|Number=Sing|Person=3 13 nsubj _ SpaceAfter=No
4 - - PUNCT Punc _ 13 punct _ SpaceAfter=No
5 dildan dil NOUN Noun Case=Abl|Number=Sing|Person=3 6 obl _ SpacesAfter=\r\n
6 achinarkan achin VERB Verb Aspect=Perf|Mood=Ind|Polarity=Pos|Tense=Pres|VerbForm=Part 13 acl _ SpacesAfter=\r\n
7 , , PUNCT Punc _ 13 punct _ SpacesAfter=\r\n
8 uni u NOUN Noun Case=Acc|Number=Sing|Person=3 11 obj _ SpacesAfter=\r\n
9 ilk ilk ADJ Adj _ 10 amod _ SpacesAfter=\r\n
10 bor bor NOUN Noun Case=Nom|Number=Sing|Person=3 11 obl _ SpacesAfter=\r\n
11 uchratgan uchrat VERB Verb Aspect=Perf|Mood=Ind|Polarity=Pos|Tense=Pres|VerbForm=Part 12 acl _ SpacesAfter=\r\n
12 paytini payt NOUN Noun Case=Acc|Number=Sing|Number[psor]=Sing|Person=3|Person[psor]=3 13 obj _ SpacesAfter=\r\n
13 esladi esla VERB Verb Aspect=Perf|Mood=Ind|Number=Sing|Person=3|Polarity=Pos|Tense=Past 0 root _ SpacesAfter=\r\n
14 . . PUNCT Punc _ 13 punct _ SpacesAfter=\r\n\r\n

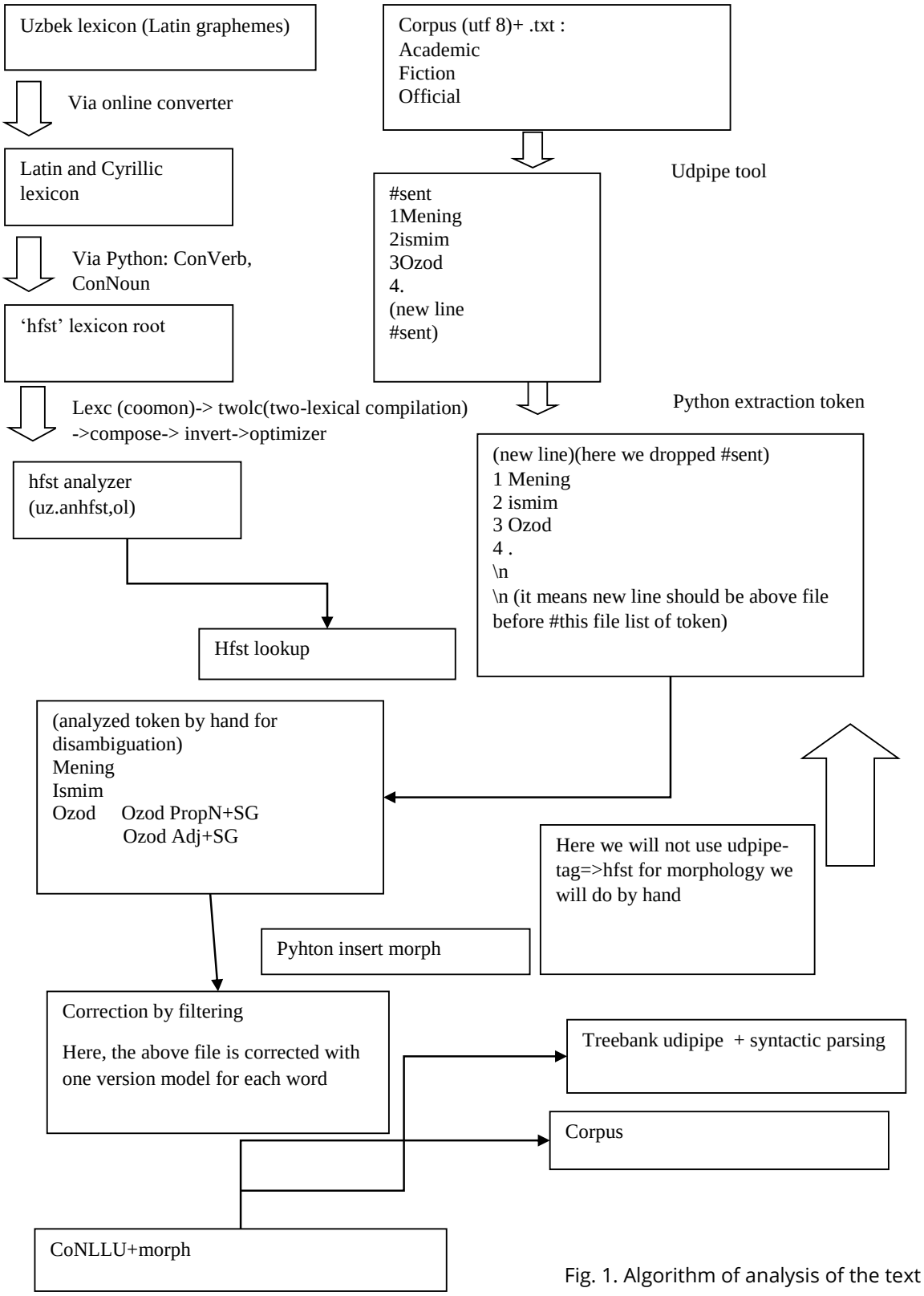
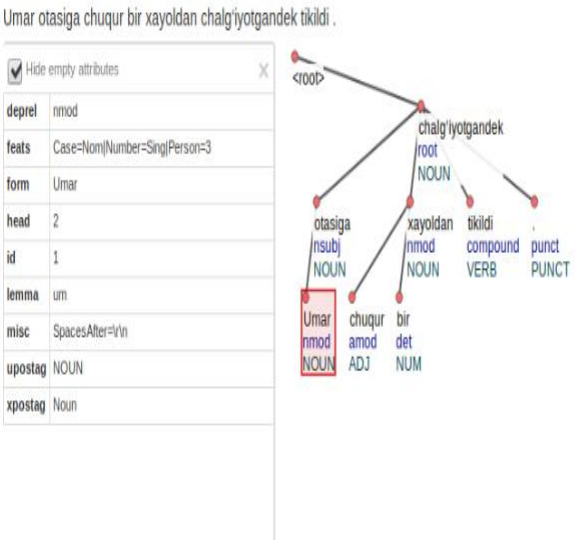


Fig. 1. Algorithm of analysis of the text

Applying CONLLU tool of universal dependency it will be represented by the following graph:



Conclusion

Universal dependency is productive tool to analyze syntactic structures of the text for relative languages. Considering the importance of syntactic parsing in corpus analysis, give good opportunity to model a number of syntactic structures of the text. One of our conclusion is manual improvement given grammatical features of each sentence of corpus can provide for disambiguation through no grammar but morphological component of parts of speech.

Acknowledgement

I would like to express my sincere gratitude to Doctor Loic Boizou for his expert advice and encouragement throughout this project by implemented "El-yurt umidi" foundation (2018, at Vytautas Magnus University in Lithuania), as well as the head of Computational linguistics center Dr. Andrius Utka for his brilliant advices.

References

Adam Przepiórkowski, Agnieszka Patejuk (2016) From Lexical Functional Grammar to enhanced Universal Dependencies The UD-LFG treebank of Polish Lang Resources & Evaluation https://doi.org/10.1007/s10579-018-9433-z

Alexandr Rosen Morphological Tags in Parallel Corpora file:///C:/Users/user/Downloads/Morphological_Tags_in_Parallel_Corpora.pdf

Carlos Gymez-Rodriguez (2010) Parsing Schemata for Practical Text Analysis Imperial College Press 5

Gülşen Cebiroğlu Eryiğit, Kemal Oflazer, Joakim Nivre (2008) Computational Linguistics · December

Abdurakhmonova N., Tuliyev U. (2018), Morphological analysis by finite state transducer for Uzbek -English machine translation / Foreign Philology: Language, Literature, Education. №3 (68), - P. 59-66

Oflazer, Kemal. (1994) Two-level description of Turkish morphology. Literary and Linguistic Computing, 9(2):137–148.

Proceedings, pages 169–172
Paris, UNESCO Headquarters, 5-6 December, 2019.

Womb Grammars: A constraint solving model for learning the grammar of Yorùbá

Ife Adebara
University of Brtish Columbia
Cognitive Systems
ife.adebara@mail.ubc.ca

Abstract

We address the problem of inducing the grammar of Yoruba from that of English. We adopt an efficient and linguistically-savy constraint solving model for an under-resourced language, Yorùbá, from the grammar of English. Our model - Womb Grammars (WG) parses a subset of noun phrases of the target language Yorùbá, from the grammar of the source language English. Our solution is straight forward and only requires a correct property grammar of the source language, a lexicon of the target language and a set of representative input phrases in the target language. This is extensible to and useful for other low resource languages where availability of large corpora is a challenge.Our proposed methodology adapts Womb Grammars in parsing phrases of the target from the grammar of English which is described as properties between constituents. Our model is implemented in CHRG (Constraint Handling Rule Grammars).

Keywords: property grammars, constraint handling rules grammar, constraint handling rules, failure-driven parsing, womb grammars, Yorùbá

Yorùbá

A fi èro Womb Grammars hàn, èyí tí ó n fa gírámà ède Yorùbá láti gírámà ède gèésí. Èro yîı yìò lo gírámà gbólóhùn óro orúko ède gèésí láti fi pín èya gbólóhùn óro orúko ède Yorùbá fún íwónba ápeere tí a fi se àlàyé isé yìí. Isé yìí ko nira fun èro Womb Grammars rárá. A kàn nílò gírámà ède gèésí èyí tí a pè ní "property grammar", ìwé ìtúmò ède Yorùbá àti ápeere gbólóhùn óro orúko ède Yorùbá. Yorùbá nìkan kó ni isé yìí wúlò fún a tún le l fún àwo ède mírà tí kò ní àkosílè ède púpò bíi ède Yorùbá.

1. Introduction

Womb Grammars (WGs) a novel grammar induction technique, developed by Dahl and Miralles (Dahl and Miralles, 2012; Becerra et al., 2013; Dahl et al., 2012). induces the grammar of a *target* language from the grammar of a *source* language. The WGs paradigm describes a language's phrases in terms of constraints or properties between pairs of direct daughters of a phrasal category called properties. WGs extends the parsing capabilities implicit in these properties into a model of grammatical induction, in addition to parsing.

In this paper, WGs are used to induce a subset of noun phrases in Yorùbá from that of English (Adebara and Dahl, 2016). We assume that the grammar of the source language is correct and that the lexicon and input phrases of the target language are correct and representative of noun phrases in the target language. We adopt a constraint satisfaction approach whereby every phrase of the target language is tested for satisfaction and unsatisfied constraints provide a lead for the reconstruction of the target grammar using the source grammar. We use a context free grammar of the target language to evaluate the correctness of our parser.

Our results in applying and adapting the WG model for inducing Yorùbá noun phrases show that this model compares favourably with others in solving the grammar induction problem: it combines linguistic formality with efficient problem solving, and can transfer into other languages, including languages in which tones have a grammatical and / or semantic function.

The rest of this paper is divided as follows: Section 2 describes the motivation for this work, Section 3 provides a description of the Yorùbá language. In Section 4, we describe the Linguistic Background of this work. We introduce the concept of property grammars and Womb Grammars in Section 5 and 6 respectively. In Section 7 we explain our results and conclude in Section 8.

2. Motivation

Language endangerment and death has been of serious concern in linguistics and language policy making. Close to seven thousand languages are currently spoken in the world, the majority of which are understudied and endangered. It has been said that an alarming 50 to 90 percent of languages will be extinct by the end of the century (Romaine, 2017).

For various reasons, some speakers of many minor, less studied languages may also learn to use a different language from their mother tongue and may even stop using their native languages. Parents may begin to use only that second language with their children and gradually the transmission of the native language to the next generation is reduced and may even cease. As a result, only the elderly in such communities may use the native language, after a while, there may be no speakers who use the language as their first or primary language and eventually the language may no longer be used at all. Thus, a language may become extinct, existing perhaps only in recordings, written records and transcription and languages which have not been adequately documented completely disappear.

Linguists cannot keep up with the study of these languages even for educational purposes, and there is a growing need for their automatic processing as well, since the amount of

text sources grows much faster than humans can process them. To make matters worse, most linguistic resources are poured into English and a handful of other first world languages, leaving the vast majority of languages and dialects under-explored. Clearly, automating the discovery of an arbitrary language's grammar model would render phenomenal service to the study and preservation of linguistic diversity.

Scientifically, we wanted to explore to what extent the parsing-as-constraint-solving paradigm of NLP problem solving could buy us a great degree of linguistic descriptive formality without sacrificing efficiency, in the realm of grammar induction and in particular for inducing Yorùbá, which is severely under-resourced and endangered.

3. The Yorùbá Language

Yorùbá belongs to the Yoruboid group of the Kwa branch of the Niger-Congo language family, which cuts across most of sub-Saharan Africa. It is a tonal dialect-continnum comprising about 20 distinctive dialects and spoken by over 30 million people in the western part of Nigeria (Gbenga, 1994). Niger-Congo is the largest of the five main language families of Africa. The others being Nilo-Saharan, Afro-Asiatic, Khoisan and Austronesian (mainly found in the nation of Madagascar).

Yorùbá is one of the three regional (national language contained in the constitution) languages in Nigeria and is said to be the most studied African language. Yorùbá is spoken by more than 20 percent of the population of Nigeria. The two other national languages are Hausa and Igbo, both of which are also regional languages in the north and south-eastern parts of the country respectively.

3.1. The Sociolinguistic Situation of Yorùbá in Nigeria

Despite the seemingly large population of Yorùbá speakers, according to (Ayo, 1993) Yorùbá has been classified as a "deprived" language. This is influenced by the language policy of Nigeria which favours the use of English above all indigenous languages. It also pays lip service to for instance the national language policy of education, which states that the mother tongue or the language of the immediate community must be adopted as the language of education in primary schools, and English should only be introduced at a later stage. Other language policies in other domains that encourage the use of mother tongues are also not adopted. So that the regional or national status of Yorùbá and other regional languages is theoretically but not fully implemented in practice (Abidemi and Segun, 2005).

English is the language of the elite and fluency in English is synonymous with a good education. As a result, many parents, even those who are barely educated or not educated at all, ensure that their children are taught in English right from the elementary classes. In most schools, indigenous languages are referred to as vernacular and are prohibited. Violation usually attract fines and many times corporal punishment. Bilingualism is also believed to affect children's ability to attain competence in English and thus parents avoid speaking mother tongues at home for fear of raising children with poor communication in English. Many children therefore many can neither speak, read nor write in Yorùbá and many do not even understand the language at all.

All the afore mentioned language situation have influenced the lack of adequate language development and consequently resulted in, resource scarcity of Yorùbá.

4. Linguistic Background

4.1. Data Collection

Data collection for this research has been a combination of introspective and empirical collection methods. This approach of data collection was employed in order to ensure that our model is realistic, correct and robust. Introspection has proven to be the most reliable process of data collection and also very useful for building models which require high level linguistic competence such as this WG model (Chomsky, 1957). Introspection has been that of the author who has formal training in linguistics and is also a native speaker of Yorùbá. We have double-checked our introspective conclusions by consulting as well seven other native speakers of Yorùbá, four of which also have formal graduate level training in linguistics.

Data collected have also been compared with two exisiting grammars of Yorùbá. The first by Ayò Bámgbósé (Ayo, 1966) and the other by Awóbùlúyí (Oladele, 1978). It was important to observe these exisiting grammatical descriptions of Yorùbá, considering that they are one of the earliest contributions of native speakers who have formal linguistic training to the description of the Yorùbá grammar.

4.2. Strategies for Part of Speech Parsing

Assigning correct part-of-speech tags to each input word explicitly indicates some inherent grammatical structure of any language and a wrong part-of-speech tag will distort the grammatical structure of a language. We adopt a rule-based (derived from linguistic rules) approach. Rule-based approach though rigorous and requiring a great amount of high level linguistic skills, yield good results for any language, including those like Yorùbá which have been identified as resource scarce as well as having a less fixed word order structure.

The tagsets were developed using the Penn Tree Bank of Yorùbá (Yiwola, 2008) as well as judgments of native speakers of Yorùbá who have formal linguistic training. The tagsets were also compared to the grammars of Ayò Bámgbósé (Ayo, 1966) and Awóbùlúyí (Oladele, 1978).

We use the following tagsets: *noun e.g ajá (dog)*, *pronoun e.g àwon (they)*, *proper-noun e.g Ayò*, *determiner e.g kan (a)*, *quantifier e.g gbogbo(every)* and adjective e.g dúdú(black). We further define features for each word in order to provide a fine-grained definition of each word tag. We *Number*, *Gender*, *Tone*, *Person*, *Definitiveness*, and *Case*. These features have been carefully chosen to ensure that our model accounts for the unique traits of Yorùbá.

5. Property Grammars

The idea of constraint is present in modern linguistic theories such as Lexical Functional grammars (LFG) and Head-

driven Phrase Structure grammars (HPSG). However, constraint satisfaction, a way of implementing constraints, is not really incorporated in the implementation of these theories. Thus, we use a formalism called Property Grammar (PG)(Blache and Rauzy, 2012) which is based completely on constraints: all linguistic information is represented as properties between pairs of constituents, which allow parsing to be implemented as a constraint satisfaction problem. For example in the PG framework, English noun phrases can be described through a few constraints such as precedence (a determiner must precede a noun, an adjective must precede a noun), uniqueness (there must be at most one determiner), exclusion (an adjective phrase must not coexist with a superlative), obligation (a noun phrase must contain the head noun), and so on. Instead of resulting in either a parse tree or in failure as traditional parsing schemes do, such frameworks characterize a sentence through the list of the constraints a phrase satisfies and the list of constraints it violates, so that even incorrect or incomplete phrases will be parsed. Moreover, it is possible to relax some of the constraints by declaring relaxation conditions in modular fashion.

6. Womb Grammars

Womb Grammar (Adebara et al., 2015; Ife and Veronica, 2015; Philippe, 2005) was presented in two versions: *Hybrid Womb Grammars*, in which the source language is an existing language for which the syntax is known, and Universal Womb Grammars, in which the source syntax is a hypothetical universal grammar of the authorsâ own devise, which contains all possible properties between pairs of constituents. We adopted the Hybrid Model.

The general WG model can be described as follows: Let L^S be the source language. Its syntactic component will be noted L^S_{syntax}. Likewise, we call the target language L^T and its lexicon (L^T_{lex}). If we can get hold of a sufficiently representative set of phrases in L^T that are known to be correct (a set where our desired subset of language is represented), we can feed these to a hybrid parser consisting of L^S_{syntax} and L^T_{lex}. This will result in some of the sentences being marked as incorrect by the parser. An analysis of the constraints these "incorrect" sentences violate can subsequently reveal how to transform L^S_{syntax} so it accepts as correct the sentences in the corpus of L_T—i.e., how to transform it into L^T_{syntax} by modifying the constraints that were violated into constraints that accept the input.

For instance, let $L^S = English$ and L^T = Yorùbá, and let us assume that English adjectives always precede the noun they modify, while in Yorùbá they always post-cede it (an oversimplification, just for illustration purposes). Thus "a red book" is correct English, whereas in Yorùbá we would more readily say "iwe pupa kan" (book, red, a).

If we plug the Yorùbá lexicon and the English syntax constraints into our WG parser, and run a representative corpus of (correct) Yorùbá noun phrases by the resulting hybrid parser, the said precedence property will be declared unsatisfied when hitting phrases such as "ìwé pupa kan". The model transformation module can then look at the entire list of unsatisfied constraints, and produce the missing syntactic component of L^T's parser by modifying the constraints in L^S_{syntax} so that none are violated by the corpus sentences.

6.1. Modified parsing that calculates both failure and success explicitly

Some of the necessary modifications are easy to identify and to perform, e.g. for accepting "ìwé pupa kan" we only need to delete the (English) precedence requirement of adjective before noun (noted $adj < n$). However, subtler modifications may be in order, after some statistical analysis in a second round of parsing: if in our L^T corpus, which we have assumed representative, *all* adjectives appear after the noun they modify, Yorùbá is sure to include the reverse precedence property as in English: $n < adj$. So in this case, not only do we need to delete $adj < n$, but we also need to add $n < adj$.

Previous models of WGs (Adebara et al., 2015; Ife and Veronica, 2015; Ife and Dahl, 2015) focused on failure driven pasing, under the assumption that failed properties are usually the complement of those satisfied, so they can be derived from the failed ones if needed and in general, a grammar is induced by repairing failures. However our more in depth analysis in the context of Yorùbá has uncovered the need for more detail than simply failing or succeeding, as in the case of conditional properties. We therefore now use a success conscious *and* failure conscious approach for inducing the grammar of our target language, Yorùbá. Each input phrase of the target language is tested with all relevant constraints for both failure and success. This makes the model slightly less efficient than if we only were to calculate failed properties, but of course the gain is in accuracy. Efficiency is still guaranteed by the normal Constraint Handling Rules Grammar (CHRG) way of operating: rules will only trigger when relevant, e.g. if a phrase is comprised of only a noun and an adjective, it will not be tested with for instance precedence(pronoun, determiner) or any other constraint whose categories are different from those of the input phrase. We keep a list of all properties that fail and another for those that succeed together with the features of the categories of each input phrase and their counts. It is important to state that constituency constraints are tested only for success. This is because we are interested in checking that our target grammar shares similar constituents with our source language and testing for failure will be irrelevant for these constraints. We also are able to induce constituents present in the target grammar that are not in the source grammar.

7. Results and supporting evidence of correctness

Our results so far have been consistent with linguistic research of Yorùbá grammar. We also use phrases generated by a Context Free Grammar (CFG) which we developed as the input phrases in our WG model and our induced grammar have shown similarities with the CFG. It is important to state that despite equivalences that our induced grammar share with the CFG subset, our induced grammar explicitly encodes more information than the CFG. This is because phrase structure representations such as context free grammars use a unique explicit relation hierarchy, that

encode constituency information so that other information such as linearity, obligatority, dependency, etc. are implicit. On the opposite, constraint-based representations, such as our property grammar encode explicitly all these relations. (Blache and Rauzy, 2012). We summarize our results into constituency, precedence, requirement, dependency and obligatority.

1 Constituency: Our constituency results show that in Yorùbá, nouns, pronouns, proper-nouns, adjectives, quantifiers and determiners are allowable categories in noun phrases as described in literature (Peter, 1970; Oladele, 1978; Ayo, 1966) as well as in our CFG.

2 Precedence: We induce two conditional precedence properties (although conditional precedence(adjective, noun) has two different conditions.), and nine precedence properties. Our conditional precedence imply that there are two orderings, in pronouns and nouns and adjectives and nouns. This we found evidence for in literature(Peter, 1970). However, we do not induce a property for quantifier and noun. This is because, there exist no known pattern responsible for the difference in order, which is consistent with research claims (Oladele, 1978).

3 Requirement: We do not induce any requirement between nouns and determiner. This is because of a lack of pattern in features where the requirement property succeeds and where it fails. This is consistent with our CFG which has rules where nouns occur without determiners as well rules where nouns occur without determiners. This conclusion is also presented in research (Ayo, 1966; Oladele, 1978; Oladiipo, 2009).

4 Dependency: Our model also does not induce dependency rules. This is because there was no unique feature present with instances where dependency failed and when it succeeded. This again is supported by (Ayo, 1966) but not explicit in our CFG.

5 Obligatority: Obligatority succeeded in all input phrases, showing that at least one of noun, pronoun and propernouns are compulsory constituents of Yorùbá. Our CFG also shows these three constituents occur at least once in all rules.

8. Conclusion

We have shown the simplicity with which Womb Grammars automatically induces the grammar of Yorùbá from that of English despite the peculiarities in the grammar of Yorùbá that can make this very difficult which makes our model very useful in language development and language documentation. Our system automatically transforms a user's syntactic description of a source language into that of a target language, of which only the lexicon and a set of representative sample phrases are known. While demonstrated specifically for English as source language and Yorùbá as target language, our implementation can accept any other pair of languages for inducing the syntactic constraints of one from that of the other, as long as their description can be done in terms of the supported constraints.

9. Bibliographical References

Abidemi, F. F. and Segun, S. A. (2005). Is yorùbá an endangered language? *Nordic Journal of African Studies*, 14(3):18–18.

Adebara, I. and Dahl, V. (2016). Grammar induction as automated transformation between constraint solving models of language. In *Proceedings of the Workshop on Knowledge-based Techniques for Problem Solving and Reasoning co-located with 25th International Joint Conference on Artificial Intelligence (IJCAI 2016), New York City, USA, July 10, 2016*.

Adebara, I., Dahl, V., and Tessaris, S. (2015). Parsing with partially known grammar. In *Agents and Artificial Intelligence - 7th International Conference, ICAART 2015, Lisbon, Portugal, January 10-12, 2015, Revised Selected Papers*, pages 334–346.

Ayo, B. (1966). A grammar of yoruba. vol. 5.

Ayo, B. (1993). Deprived, endangered, and dying languages. *Diogenes*, 41(161):19–25.

Becerra, L., Dahl, V., and Miralles, E. (2013). On second language tutoring through womb grammars. Accepted for publication at IWANN 2013, June 12-14, Tenerife, Spain.

Blache, P. and Rauzy, S. (2012). Hybridization and treebank enrichment with constraint-based representations. In *Proceedings of LREC*.

Chomsky, N. (1957). Syntactic structures [text]. In *Walter de Gruyter*, page 117.

Dahl, V. and Miralles, J. E. (2012). Womb grammars: Constraint solving for grammar induction. In J. Sneyers et al., editors, *Proceedings of the 9th Workshop on Constraint Handling Rules*, volume Technical Report CW 624, pages 32–40, Department of Computer Science, K.U. Leuven.

Dahl, V., Miralles, E., and Becerra, L. (2012). On language acquisition through womb grammars. In *7th International Workshop on Constraint Solving and Language Processing*, pages 99–105.

Gbenga, F. J. (1994). *The Yoruba Koiné: Its History and Linguistic Innovations*, volume 6. Lincom Europa.

Ife, A. and Dahl, V. (2015). Domes as a prodigal shape in synthesis-enhanced parsers.

Ife, A. and Veronica, D. (2015). Shape analysis as an aid for grammar induction.

Oladele, A. (1978). *Essentials of Yoruba grammar*. University Press Plc Nigeria.

Oladiipo, A. (2009). Analyzing yoruba bare nouns as dp. *Lagos Notes and Records*, 15(1):30–55.

Peter, O. (1970). The essentials of the yoruba language.

Philippe, B. (2005). Property grammars: A fully constraint-based theory. In *Proceedings of the First International Conference on Constraint Solving and Language Processing*, CSLP'04, pages 1–16, Berlin, Heidelberg. Springer-Verlag.

Romaine, S. (2017). The impact of language policy on endangered languages. In *Democracy and human rights in multicultural societies*, pages 217–236. Routledge.

Yiwola, A. (2008). Global yoruba lexical database.

Proceedings, pages 173–176
Paris, UNESCO Headquarters, 5-6 December, 2019.

The Contribution of StoryWeaver in Keeping Indigenous Languages Alive

Archana Nambiar
Research Consultant, Pratham Books
#621, II Floor, 5th Main, OMBR Layout, Bhuvanagiri Main Road, Lakshmamma Layout, Banaswadi, Bengaluru, India 560076
archana@prathambooks.org

Abstract

StoryWeaver, an open access digital content platform from Pratham Books has catalyzed the creation of quality reading materials for children in indigenous and mother tongues. These reading materials are useful resources for literacy development and serve as the much needed link between the home language and the medium of instruction in school. The platform's open licensing model has synergized efforts of indigenous communities, linguists, language enthusiasts, educationists and translators in creating resources in indigenous languages hitherto underrepresented in the mainstream. StoryWeaver's innovative model of combining technology, open licensing and collaboration has contributed to preserving the linguistic diversity of indigenous communities.

Keywords: StoryWeaver, indigenous languages, children's literature

Résumé

प्रथम बुक्स के ऑनलाइन पठन सामाग्री मंच स्टोरीवीवर ने देशीय और मातृभाषाओं में बच्चों के लिए गुणवत्तापूर्ण सामग्री का सृजन किया है। कहानी पुस्तक के रूप में इन पठन सामग्रियों ने साक्षरता को बढ़ावा देते हुए, पठन आनंद को बढ़ाकर घरेलू और विधिवत शिक्षा की भाषाओं के बीच की कड़ी का काम किया है। प्लैटफॉर्म के ओपन लाइसेंसिंग मॉडल ने मुख्यधारा में आने वाली देशी भाषाओं में कहानियाँ रचने और प्रकाशित करने के विभिन्न भाषाई समुदायों, भाषाविदों, भाषा शिक्षाविदों और अनुवादकों के कार्य को गति दी है। स्टोरीवीवर की साझेदारी ने कई समुदायों की भाषाई विविधता को संरक्षित करने में योगदान दिया है साथ ही दूरवर्ती क्षेत्रों में भी बच्चों की साक्षरता को बढ़ावा दिया है।

Introduction

Languages are intrinsically linked to culture and central to the identity of communities. They are of strategic importance for people with implications for identity, communication, social integration and development. Language supports cognitive processes required for learning. It is estimated that 96 per cent of the world's approximately 6,700 languages are spoken by only 3% of the world's population. Even as indigenous peoples constitute less than 6% of the global population, they speak more than 4,000 of the world's languages. It is believed that more than half of the world's languages will become extinct by 2100 (The United Nations Permanent Forum on Indigenous Issues, 2018).

For indigenous communities, languages are central to their identity, the preservation of their cultures, worldviews, visions and an expression of self-determination. With globalization and the rise of a small number of culturally dominant languages, indigenous languages are increasingly under threat and no longer transmitted by parents to their children (ibid). While a few hundred languages have a legitimate presence in the public domain and the digital world, a majority of the languages are fading without being given any space for exposure, growth and nurturance.

International agencies have drawn attention to the critical loss of indigenous languages and the urgent need to protect, revitalize and promote linguistic, cultural diversity and multilingualism.

StoryWeaver (https://storyweaver.org.in/) is an online content platform from Pratham Books that has its roots in promoting access to reading materials for children. Pratham Books is a not-for-profit children's book publisher based in India that seeks to address the critical shortage of reading materials for children in mother tongues by publishing high quality, low cost storybooks to promote reading acquisition. StoryWeaver was launched with the idea that by openly licensing a large repository of children's storybooks and by providing simple digital tools, it would allow users to translate or version the books into new languages and forms, and enable the creation of more books, thereby having greater scale and impact. All content on the platform is available under the Creative Commons license – CC BY 4.0 that allows use, reuse and remixing of the material. The platform can be accessed on desktops, laptops, and mobile devices and its interface is simple, intuitive, and non-bandwidth intensive. All the content on StoryWeaver is Unicode compliant which makes it discoverable, accessible and interoperable.

StoryWeaver has Helped Address the Dearth of Children's Literature in Indigenous Languages

Education through the medium of a dominant language creates pedagogic and psychological barriers for children from indigenous communities. Often dominant languages are learned by children at the cost of displacing the mother tongue. As a result, children's linguistic repertoire in the mother tongues diminishes as they take on the linguistic and cultural traditions of the dominant language, contributing to the disappearance of linguistic diversity. The education system at large is ill-equipped to use multilingualism as a resource. Further, the lack of access to a wide variety of children's books in mother tongues severely limits meaningful participation of

children, communication and expression.

StoryWeaver has catalyzed the creation of quality reading materials for children in indigenous languages that are largely ignored by mainstream publishers. StoryWeaver's innovative model has as its core a large repository of high quality, openly licensed storybooks from publishers such as Pratham Books, Book Dash, Room to Read and others, which can be read online, offline or printed and used. Users can also translate and the original content into new languages and this has led to a huge amplification of books for children especially in minority and indigenous languages. The model is predicated on the fact that there is a huge shortage of books in mother tongue languages for children and by empowering local actors to participate in the process of translation of original content, the book gap can be narrowed. In just 4 years, StoryWeaver has grown from a repository of 800 storybooks in 24 languages to over 17,000 storybooks in 208 languages, of which 40% are indigenous languages. The books have been read over 11 million times online and offline and the platform is used by 2.2 million users in 150 countries. Every year, StoryWeaver marks International Mother Language Day with Freedom to Read, a campaign to build open hyperlocal digital libraries in mother tongues. Collaborations with a global network of organizations and individuals have ensured the development of reviewed, high-quality books in indigenous languages that were hitherto inadequate or non-existent. Each of these hyperlocal digital libraries encompasses 50 to 100 storybooks and is continually expanding.

Suchana, a non-governmental community organization based in Birbhum district of West Bengal, India has been promoting learning among tribal children through their mother tongues in the early years. Both Santali and Kora have a rich oral tradition but no children's books. Teachers who teach children who speak Santali and Kora are usually completely bereft of resources for early literacy development. Prior to 2014 and the launch of StoryWeaver, they had developed 15 educational resources in these languages. The collaboration with StoryWeaver allowed them to translate 105 books into Santali and 100 into Kora within a short span of two years. All these storybooks are published on StoryWeaver and can be used by others as well. 10,000 copies of 20 titles have been printed and distributed to government schools, pre-schools and other organizations in the region. Both the print and digital books are incorporated into their mobile library programme which reaches around 3000 children in 25 villages and formal schools and pre-schools. These efforts have resulted in children having access to a range of books in their languages, including bilingual ones that help them to transition from their mother tongues into Bengali, the state language.

QwertyWorks, a for-profit translation enterprise based in Manila, Philippines has been translating and distributing books in the local community centres since 2001. Most storybooks, both for adults and children in the country are in the official language, English. The few books that are available in native languages are not affordable. Most children go through the schooling process without ever having read books in their mother tongues. Although Cebuano speakers constitute about one-fifth of the population of the Philippines, it is little used as a literary language, with a dearth of children's literature. Through QwertyWorks' partnership with StoryWeaver, 43 books were translated into Cebuano-Cebu. The books were also distributed through Libreo.ph, an online portal and Facebook. During the pilot project, the books were digitally printed and distributed to 30 day care centers benefitting around 1200 children. These were the first ever books that these children had access to in their own language.

Korku is a tribal language spoken in the regions of Vidarbha in Maharashtra and Khandwa in Madhya Pradesh, India by nearly 200,000 people. Unnati Institute for Social and Educational Change (Unnati ISEC) has been working in the district of Akola, Maharashtra with children from the Korku community for the last five years. Korku children face a tremendous language disadvantage in schools where the medium of instruction is Marathi. Unnati ISEC is working to foster an inclusive culture within the school where the child's language and culture are reflected in the curriculum, books and pedagogy. Prior to their collaboration with StoryWeaver, they had published around 100 books in Korku and Marathi in a very time consuming manner. The partnership with StoryWeaver added 100 new storybooks in Korku. Printouts of the storybooks are used in libraries and classrooms as these schools do not have the necessary infrastructure to read books digitally. The books have certainly created the joy of reading for children who were otherwise deprived of reading material in their mother tongue.

Pawari is a tribal language spoken in Maharashtra with around 360,000 speakers. Children who speak Pawari learn in Marathi, which is the medium of instruction and often find it difficult to understand the concepts being taught. An educator, Amit Dudave has been working towards making storybooks available in Pawari. He has translated 26 storybooks into Pawari which are being used in schools in the region.

StoryWeaver has Facilitated the Revitalization of Indigenous Languages and Building of a Strong Cultural Identity for Participation in Cultural and Social Processes

For indigenous peoples whose languages are steeped in oral history and storytelling since time immemorial, it forms the basis of their cultural identity. There is a growing recognition among indigenous communities, practitioners and policy makers for the need for revitalizing indigenous languages.

In most African countries, indigenous languages are relegated to home and local interactions. The mediums of instruction at school remain the dominant or official languages. Although children have access to books in schools, there are very few high quality reading materials in the indigenous languages. The African Libraries and Information Associations and Institutions (AfLIA) pursues the interests of library and information services in 28 countries in Africa and have the promotion of the mother tongue as a primary agenda. As part of StoryWeaver's Freedom to Read campaign, AfLIA organized a series of translation workshops, where more than 200 books were translated in Ewe, Fante, Hausa, Igbo, Isixhosa, Kikuyu, Luganda, Swahili and Yoruba. The translated books are being used in the continent-wide reading promotion 'Read Africa Read' where the same books are read in different languages across different

countries. Individual libraries have been using the storybooks for their story hours and literacy sessions for out-of-school children and adults.

Chinyanja/Chichewa which is widely spoken in Malawi, Zambia and Mozambique lacks children's books. Writer, translator, editor and poet, Agnes N.S. Nyendwa has been translating books into her mother tongue, Chinyanja. Most children who speak Chinyanja at home do not have adequate reading resources that help them develop fluency in the language. 45 titles from StoryWeaver have been translated into this language.

S'gaw Karen is spoken by over 200,000 people in Thailand. There is a scarcity of books in this language, which is predominantly an oral language. The Asia Foundation is an international non-profit that helps societies work towards a peaceful, just, and thriving region. Their focus has been on creating quality reading resources in indigenous languages in Asia through technology initiatives, translations and content creation. Their Let's Read! digital platform is designed to create and adapt open-access books into any number of languages, quickly increasing the number of relevant and high-quality children's storybooks available. Through community workshops, publishers, authors, editors and technologists have come together to create engaging content in national and indigenous languages in digital format. There are about 256 different versions of StoryWeaver books available in 9 languages on the Let's Read! platform. The books are freely available for anyone to use on the platform, an Android reader application, or to be printed for non-digital use.

In an effort to give a voice to tribal cultures that do not find adequate representation in mainstream discourses, Pratham Books partnered with Ignus Educational Resource Guild (Ignus ERG), an educational resource group to create the Adikahani series of storybooks. A set of 10 tribal stories and 4 song cards authored and illustrated by tribal communities in four tribal languages spoken in Odisha, India– Juanga, Kui, Munda and Saura were developed. These storybooks were published on StoryWeaver and are being used in schools where Munda and Kui children study.

Surjapuri is a minor language spoken in pockets of Bihar, West Bengal, Assam in India and Bangladesh by 1.2 million people. Unfortunately, there are no books in this language and a real possibility of the language being lost as most people speak Hindi. The Azad India Foundation (AIF) has been working towards improving the health and educational status of marginalized women, adolescents and underserved children of Kishanganj district in Bihar. A total of 100 books developed by AIF are now being used in the 75 learning centers run by the organization. Creating a hyperlocal library on StoryWeaver has helped the children have access to and preserve Surjapuri as their language.

Toto is an oral language with no script and spoken only in a village called Totopara in Alipurdur in West Bengal, India. Under a project called the Study and Research of Indigenous and Endangered languages of India of the Jadavpur University, efforts were made to create a storybook on StoryWeaver to document the language.

StoryWeaver has Supported the Creation of Local Publishing Models for Indigenous Children's Literature

The publishing space for indigenous literature in general and children's literature in particular has been diminishing. For the most part, indigenous publishing across countries remain isolated, underfunded and away from the mainstream marketplace. StoryWeaver has opened up the possibilities of supporting mechanisms for communities to showcase indigenous writing and translations, and fill a void in materials for children, and to further the development of indigenous children's literature. The development of the translation based publishing practices has been fuelled by the increasing demand for authentic storybooks for children in their languages. Several innovative publishing models have emerged as a result of collaborations with StoryWeaver.

StoryWeaver's translation sprint process accelerates the creation of local language reading resources for children. It involves a two day workshop that brings together translators to translate, review and digitally publish storybooks for children in a chosen language or languages on StoryWeaver.

Gondi is an indigenous language spoken by a tribal population of over 2.7 million adivasis across seven states in India. It is marked by an absence of a standard lexicon and is heavily influenced by the state language. Since 2014, groups working with Gondi communities have come together to preserve their language and culture. With support from Microsoft Research India and CGNet Swara, they were exploring the possibilities of taking the language online. In September 2018, in StoryWeaver's Gondi translation workshop, StoryWeaver worked with 20 Gondi speakers from three states and translated storybooks from Hindi to Gondi. Around 250 storybooks were translated and contributed towards building a repository of books for future generations of children from the Gond community. Additionally, the translated text also served as inputs for building a Gondi machine translation tool. Through volunteer efforts several of these books have been printed and distributed to children in Bastar, Chattisgarh.

The Sub-Saharan Publishers is an indigenous Ghanaian publisher that specializes in African literature and children's books. They have published three storybooks from their popular Fati series on StoryWeaver in English and northern Ghanaian languages, namely, Dagaare, Dagbani and Sisali. The universal appeal of the stories has resulted in translations in multiple African languages and non-African languages.

Although Uganda has over 65 indigenous languages, none of these languages are official except for English and Kiswahili. There are very few books available in the indigenous languages for both instruction and pleasure reading. The Uganda Christian University (UCU) aims to support early literacy initiatives by making quality multilingual storybooks freely available to the local communities. Students of the creative writing course at UCU have created, translated and published 340 storybooks in 15 languages on StoryWeaver. UCU has facilitated the printing of selected titles and parents, authors and faculty members are encouraged to share these stories on Facebook, Whatsapp, downloads on mobile phones and tablets. So far 3000 copies of storybooks have been printed for distribution to community libraries, schools and children. As a result of the collaboration, storybooks are now available in

languages such as Acholi, Ateso, Kumam, Kinyarwanda, Lango, Luganda, Lumasaaba, Lusoga, Rufumbira and Runyoro.

Numerous indigenous languages in Mexico are on the verge of disappearance. In most indigenous communities, children no longer speak their mother tongue. Bilingual education is not effectively practiced in educational settings due to lack of appropriate training of teachers and resources. Alfredo Harp Helu Foundation has been supporting initiatives to promote language preservation and culture in Mexico. Using the StoryWeaver platform, 122 storybooks from Spanish have been translated to 11 indigenous languages such as Chatino, Chocholteco, Huave, Mixe, Mixteco, Nahuatl, Popti', Tarahumara, Triqui, Zapoteco and Zoque. Storybooks in Chocholteco, a highly endangered language with only 200 speakers are the first written texts in the language in over a decade. Translations in these languages have been carried out by teachers, linguists and translators and are available on the website www.colmix.org. Efforts are on to integrate the storybooks with the Endless OS, an open source operating system provided on low cost computers to schools. The Chatino and Mixe books are being used in the printed form to promote reading among children.

Conclusion

StoryWeaver's model of collaborative content creation which relies on a core of openly licensed quality content and uses technology to facilitate the process for translations is contributing to preserving the linguistic diversity of indigenous communities. Its open frameworks have synergized the efforts of civil society groups, linguists, language enthusiasts, educationists and translators in documenting and creating storybooks in languages hitherto underrepresented in the mainstream. Storybooks developed and published on StoryWeaver have been the first children's books published in some of these indigenous languages. In some cases, these have been the first ever reading material published in the indigenous language. Several of the indigenous languages that are on the platform have a digital presence for the first time as well.

The translation sprint model has been optimally used to create hyperlocal libraries in indigenous languages. StoryWeaver partners have worked around to address the challenges related to using a script and appropriate expression in the indigenous languages. An important spin off has been the development of capacities of indigenous community members to become authors and publishers capable of reclaiming their language.

StoryWeaver has facilitated the development of early reading bilingual resources in several tribal languages to express tribal identities and aid transition of the tribal child towards the language of instruction. These storybooks have facilitated children to have a deeper emotional connect with the material leading to more authentic learning experiences. The storybooks with their simple, creative and child friendly content have proven to be useful resources in early grade classrooms. Teachers, both indigenous and non-indigenous, find these books contextual, culturally sensitive and appropriate for teaching early grades.

Moreover, indigenous communities feel that the availability of content in their mother tongue both in digital and print formats has given their language a sense of legitimacy and pride. While the majority of the storybooks developed by the partners have been translations, the platform has given impetus to create books from and about the community as well. The unprecedented scale in the access to material in one's mother language realized through StoryWeaver will continue to inspire communities to access and create relevant content for children in the years to come.

Acknowledgements

We thank African Libraries and Information Associations and Institutions, Agnes N.S. Nyendwa, Alfredo Harp Helu Foundation, Amit Dudave, Azad India Foundation, CGNet Swara, Ignus ERG, Jadavpur University, QwertyWorks, Suchana, The Asia Foundation, The Sub-Saharan Publishers, Uganda Christian University and Unnati ISEC for their inputs for this paper.

Bibliographical References

The United Nations Permanent Forum on Indigenous Issues, Backgrounder-Indigenous Languages, 2018

Proceedings, pages 177–181
Paris, UNESCO Headquarters, 5-6 December, 2019.

Can we Defuse the Digital Timebomb? Linguistics, Speech Technology and the Irish Language Community

Ailbhe Ní Chasaide, Neasa Ní Chiaráin, Harald Berthelsen, Christoph Wendler, Andrew Murphy, Emily Barnes, Christer Gobl
Trinity College, Dublin, Ireland
anichsid@tcd.ie, neasa.nichiarain@tcd.ie
www.abair.ie

Abstract

Does speech/language technology represent a 'digital timebomb' - or an unprecedented opportunity - for minority and indigenous languages? For successful outcomes, technology development must address linguistic challenges, answer to the needs of the local language communities, enlisting them as a central partner in development. The Irish language ABAIR initiative is building (i) linguistic resources, (ii) core technologies, and (iii) applications for public, educational and access/disability use. The Government's *Digital Plan for Irish Speech and Language Technology* provides a model of the support needed by minority languages in the digital age, if the language is to feature in everyday community activities.

Keywords: Irish, speech technology, digital strategy, language community, education, disability

Résumé

An dainséar nó deis dúinn an réabhlóid dhigiteach? Do mhionteangacha agus do theangacha i mbaol, caithfidh forbairt na teicneolaíochta oiriúint do struchtúr na teanga agus do riachtanais an phobail. Sa tionscadal ABAIR a dhíríonn ar theicneolaíocht na hurlabhra, tá trí réimse taighde ar siúl: (i) bun-acmhainní teangeolaíochta, (ii) teicneolaíochtaí, agus (iii) áiseanna don phobal, do lucht an oideachais agus dóibh siúd faoi mhíchumas. Tá *Plean Digiteach do Theicneolaíocht Urlabhra agus Teanga na Gaeilge* an Rialtais mar mhúnla don tacaíocht atá de dhíth ar mhionteangacha sa ré dhigiteach. Beidh toradh na hoibre ag brath ar chomhpháirtíocht idir theangeolaithe, theicneolaithe agus phobal na Gaeilge.

1. Timebomb or Unprecedented Opportunity ?

Speech and language technologies have become deeply embedded in our daily lives and are increasingly central to how we work, interact socially, access information and education. For a language that is not digitally available the language community is forced to switch to the major language for access to these resources. Increasingly the technology is seen as a 'digital timebomb' (Evans, 2018), which is forcing the shift from the indigenous to the major language (Mac Thomáis, 2018).

Precisely because of its increasingly central presence in our lives, digital technology also presents an unprecedented opportunity for the documentation, preservation and revitalisation of the endangered language. There is a growing awareness that the provision of speech and language technologies is one vital strand in ensuring language survival. The Irish Government is launching a *Digital Plan for Irish Speech and Language Technology*, which will provide a 10-year roadmap for research and development (see Section 4).

Irish is an endangered language (Moseley, 2013), spoken as a community language in small '*Gaeltacht'* areas, mostly in the West of Ireland. Nonetheless, as an official language of the State, and since 2007 of the EU, it enjoys considerable State support and is a core curricular subject in Irish primary and second level schools.

In this paper, we describe the experience of the Irish language initiative ABAIR, which is developing speech and language technology for Irish (ABAIR, 2019). This entails a longterm strategy of (i) basic linguistic research, (ii) core technology building, and (iii) developing applications which draw both on the linguistic resources of (i) and the technologies of (ii). The development of applications answers to the needs of the language community. The focus has been on providing not only public resources but also on tools for education and disability/access.

While technology development in major languages is typically driven by commercial considerations, for the minority and indigenous language it is important to focus on the community's own priorities. This can mean a different trajectory: for example, recognition and synthesis in call centres to handle large volumes of calls are hardly necessary, but their use in the teaching of the language can have far reaching implications in the minority language context.

2. Linguistic and Sociolinguistic Challenges

Each language presents its own complex linguistic and sociolinguistic set of challenges. Irish has a number of structural features that make it different from English and many of the European languages. It is a verb-initial language (VSO) ; it is highly inflected with alternations affecting both the beginnings and ends of words (see Figure 1a for an illustration of inflected forms for the word *bád* /bˠ aː d̪ˠ/ 'boat'). A striking feature of the language is its sound system, where there is a contrast of palatalized and velarized consonants (see Figure 1b) (Ní Chasaide, 1999). Simply put, there are twice as many consonants as in, for example, English. This quality distinction among consonants serves not only to differentiate words, but is intricately linked to the system of grammatical inflection. For example, switching

between a final velarized and palatalized consonant differentiates between the nominative and genitive case, as in *bád* /bˠ a: d̪ˠ/ vs *báid* / bˠ a: d̪ʲ/ (see first two items in Figure 1a). As the Roman alphabet does not provide different letters for this consonantal distinction, the orthographic conventions for signaling consonantal quality entail the use of adjacent vowel letters. The vowel letter closest to the consonant (before or after) indicates it as belonging to the palatalized or velarized set. This, of course, results in complex strings of vowels in orthographic forms. As the illustration in Figure 1c shows, many strings of vowels correspond to a single vowel phoneme (in this case /i:/) as many of these vowel letters are simply markers of consonantal quality. Structural aspects of the language such as these are important considerations in developing resources and technologies and they are crucial to many of the applications, particularly those targeting language teaching/learning.

As in many minority languages, there is no spoken standard but rather strikingly different dialects. This has implications for technology development. In the case of Irish, it was necessary to consider multidialect provision from the outset, as this would be essential for its acceptability, uptake and usage by the community. The importance of this factor is not necessarily immediately obvious to an outside agency.

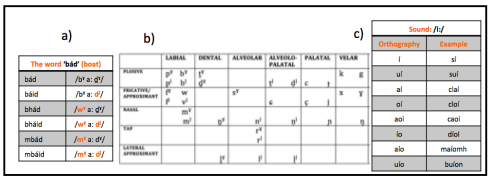


Figure 1a, 1b, 1c: Features of Irish phonology, morphology & orthography

As mentioned, although Irish survives as a community language in small, geographically separated and rather isolated areas, the language is taught nationally to school going children and adolescents. This, however, presents many challenges. Many structural aspects of the language, such as those above, are generally poorly understood by learners or teachers. Given the interconnectedness of the sound system and the grammar, and the complex writing conventions, learner outcomes in terms of pronunciation, grammatical accuracy and literacy development are often disappointing. Compounding these difficulties is the fact that learners have limited access to native speaker models of the language. Most teachers are second language learners who may themselves have an incomplete model of the language. Teaching materials tend to be dated and relatively unattractive compared to what is available for other languages.

The absence of speech and language technologies has had a particular impact on those with disabilities, excluding them from full participation in Irish language education and social (online) activities.

The lack of Irish digital technology impacts particularly on the younger digital generation, on whom the future of the language depends. It contributes to an outlook where the language is associated with the past, with poverty, with all that is not modern, alienating the younger generation who may not see the value of the language as it is disassociated from many of their daily activities, their future work and social aspirations.

3. Three Strands of Development

ABAIR entails three parallel strands of research and development, which can be illustrated in terms of the tree in Figure 2. The roots of the work entail basic research to provide linguistic components, i.e. the digitally available resources that are needed. The roots feed the trunk of the tree, i.e. the core speech technology systems. To date, speech synthesis has been the primary focus, but this is now extending to recognition and dialogue systems.

The fruits of the research are the user applications, from which the public impact of the technology will mostly be felt. These applications draw on the core technologies as well as on the linguistic components and target the general public, educational needs and the inclusion of those with disabilities. The circle of people surrounding the tree in Figure 2 represents the language community, the raison d'être for the work and its guiding force.

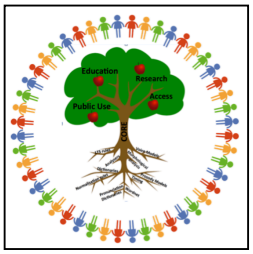


Figure 2: The 3 strands of the ABAIR initiative

3.1 Linguistic Resources

The linguistic resources include the design and development of phonetically annotated corpora. These need to be phonetically balanced (including coverage of every permutation and combination of consonants and vowels). They need to be provided for each dialect and based on dialect-appropriate materials. Additionally, letter-to-sound rules and lexica are needed for each dialect, as well as other resources such as prosody models.

Linguistic research is being carried out on many aspects of the language that had hitherto not been documented. For example, there is considerable ongoing work on the prosodic system of Irish (Dalton and Ní Chasaide, 2005; Dorn and Ní Chasaide, 2015; O'Reilly and Ní Chasaide, 2016) revealing major crossdialect differences, which our technology must capture. Considerable research effort has also gone into the creation of dialect-appropriate corpora, lexica, etc.

3.2 Core Technologies

3.2.1 Multidialect synthesis

TTS voices in three dialects have been developed and are freely available to the public at www.abair.ie. The most recent voices have deployed DNN approaches but HTS and unit-selection voices are also available.

The website is widely used : in the first three months of 2019 it had 155K hits, with many of these being from

outside Ireland. This brings home the potential role of digital technology, revealing a global community we were hitherto unaware of, and creating broader networks which are a powerful source of strength and future support.
Current work on synthesis is extending coverage from the main dialects to those lesser spoken dialects to ensure they are not left behind by these technologies. A synthetic voice for the local dialect preserves a virtual speaker: a valuable resource for the local community trying to hold on to its language, now and for future generations.
Provision of child as well as male and female 'speakers' is a further priority. Furthermore, given the bilingual context in which Irish synthesis is used, and given the frequent code switching in Irish (a feature of all minority languages), work is underway to provide bilingual Irish+English voices (e.g. for bilingual websites) that can also provide codeswitching.
Dialogue systems and interactive games are future priorities (see below). Therefore, ongoing research is modelling the prosodic/voice quality modulations needed to capture the affective nuancing such applications require (Murphy et al. 2019 ; Yanushevskya et al. 2017).

3.2.2 Current Early Stage Developments

Development of speech recognition is underway and a preliminary system has been created. Success in this area will depend crucially on recordings from a very large number of speakers. To be useful in future applications, recognition for Irish must accomodate dialect diversity, native speakers/learners, children/adults. A targeted crowdsourcing initiative *Míle Glór* (' a thousand voices') has been launched, which allows Irish speakers to record as many prompts as possible from materials that are appropriate to the speakers' dialect. It identifies categories of speaker (child/adult, specific dialect etc).
Early research towards spoken dialogue systems is also underway and will progress as speech recognition becomes available. A pilot intelligent tutoring system, *Taidhgín* has been developed. Here the learner input is through text, which is then spoken aloud by a TTS voice, and spoken answers are delivered by a talking monkey. Evaluations in schools yielded very positive reactions and dialogue systems are a priority for future research.

3.3 Applications

As mentioned, applications target the broader public as well as specific education and disability/access needs.

3.3.1 Applications for the Language Community

The public webpage at www.abair.ie, has of itself become a public application. Parents use it to help their children with homework (being able to access the pronunciation of written words helps, overcome the obstacle of the complex mapping of orthography → sounds). Foreign users use it for this and wide diversity of reasons: e.g. to pronounce Irish names, to get authentic pronunciation of Irish utterances in a play, etc. An app version of the webpage is being launched, making public access simpler. An add-on feature to web browsers allows all online text to be read aloud. This gives ready access to the spoken realization of dictionary content, online newspapers, email, etc. It also allows the Government to fulfill their statutory obligation to make information and webpages accessible to all, through the medium of Irish.

Figure 3: Homepage of the ABAIR.ie initiative

3.3.2 Educational Applications

The future of the language depends overwhelmingly on how successfully we can transmit it to the younger generation. Irish language teaching and learning can potentially be transformed by digital speech and language technology and much of ABAIR's activity to date has explored how best to deploy all resources as they come onstream. Combining the core technology of (ii) and the linguistic resources of (i) presents unprecedented opportunities for linguistically informed content and technologically advanced platforms that are pedagogically effective, attractive and motivating for the language learner. Two of the educational applications piloted to date are reviewed briefly here, targeting early and more advanced/older learners respectively. These applications bring native speaker models of the language into the classroom, making the spoken language central for all aspects of language learning. They also put the power of multimodal games and digital technologies at the disposal of learners and teachers.
The game '*Lón don Leon*' ('lunch for the lion') aims to develop phonological awareness (the ability to perceive and produce the palatalized/velarized distinction) and early literacy skills (an explicit grasp of the phonic basis for the spelling rules of the language, where vowel letters sometimes mark consonantal quality and sometimes represent actual vowel targets). Phonological awareness is trained through the use of minimal pairs, where the palatalized/velarized consonant contrast is essential to the differentiation of two words, here 'lunch' and 'lion'. This is done by embedding them in newly composed songs, illustrations and stories – reinforcing learners' awareness of the contrast and their ability to reproduce it correctly. Through games and language learning activities the learners' acquisition is monitored. When the sound contrasts are adequately acquired, the orthographic letters are introduced. Differential colour coding of palatalized (orange) and velarized (blue) versions of consonant and vowel letters guide the young learner to an understanding of how they combine in the writing system. This game can be used in the home or in a classroom setting.

Figure 4: *Lón don Leon* phonological awareness game.

A second platform, *An Scéalaí*, 'the Storyteller', targets

the older learner and has as an objective the parallel training of all four language skills: writing, listening, reading and speaking (Ní Chiaráin & Ní Chasaide, 2018; 2020). The cover page of *An Scéalaí* is illustrated in Figure 5. On the face of it this platform presents as a writing tool, which includes spelling and grammar checkers. However, the TTS voice output is a central feature and the learner is led to have all written content spoken aloud. Prooflistening is used as the initial strategy for learners to correct their own written work. Here, constant reinforcement of the links of sound → orthographic sequence is essential and evaluations to date show it to be a highly effective strategy in developing learners' awareness of many types of errors. As an intelligent Computer-Assisted Language Learning (iCALL) application, there are many types of inbuilt prompts to review and correct written work. Many draw on the specific linguistic resources developed in the first research strand above. The prompting is carried out by an interactive dialogue partner who both speaks to the learner (using their preferred TTS voice/dialect) and provides spoken feedback along with written guidance. The dialogue partner can detect areas of recurring weakness, e.g. irregular verb conjugation, and offer the learner opportunities to carry out exercises targeting these. This platform is entirely suited to autonomous learning but is also set up to allow its integration in the classroom where the teacher can monitor their own students' progress and provide further oversight and feedback. This platform is currently under evaluation in second and third level educational settings in Ireland and abroad.

Figure 5: iCALL platform *An Scéalaí* ('the Storyteller').

3.3.3 Disability/Access Applications

These technologies are particularly important for the inclusion of those with disabilities. Urgent requests from parents of schoolgoing children with visual disabilities led to the development of screenreading software, using the open source NVDA framework (McGuirk, 2015). Users control the spoken output in terms of speed and dialect, and simultaneous Braille output is provided (see Figure 6 below). Multimodal textbooks for the visually impaired have also been developed (Ní Chasaide et al. 2019).

Figure 6: visually impaired user listening & using the Liblouis Braille system simultaneously.

Current educationally-oriented multimodal platforms and games are also intended to provide for the assessment of phonological awareness and reading skills and eventually as the basis for literacy remediation for those with dyslexia (Barnes et al. 2018). A 'design for all' approach for ABAIR applications aspires to include the widest possible groups of users, especially those with disabilities.

4. The Digital Plan for Irish

The Department of Culture, Heritage and the Gaeltacht is launching a *Digital Plan for Irish Speech and Language Technology*, which will seek to ensure that the power of digital technology is available to the Irish language community. It encompasses not only the development of specific core technologies (recognition, synthesis, machine translation, etc.) but also the range of applications and linguistic research that will enable the Plan to have a longterm impact. It also aims to set in place the infrastructure to ensure the future capacity to keep up with the rapid evolution of these technologies.

5. Central Role of Language Community

The extent to which digital technology serves to revitalize the language depends on the extent to which the language community make it their own. The ABAIR experience demonstrates how, over the years, collaboration with the language community has moved from informant → requesters of applications → partners in design, testing and dissemination of specific user applications. The *Digital Plan* acknowledges the importance of the native speaker Gaeltacht community in this entreprise, as the repository of living language 'experts' and seeks its engagement in every aspect. It further aims to ensure that native speakers are among the leading researchers in future digital development. It also seeks to engage in collaborative ventures with the broader Irish language community and user groups at home and abroad.

6. Conclusion : a Model for other Minority/Indigenous Languages

Although the Irish language is endangered and the Gaeltacht communities fragile, Irish is in a relatively privileged position of being a national language with EU recognition. Its survival to date owes much to State support. The work of ABAIR has only been possible due to State funding and enthusiastic support.

It is a basic principle of the ABAIR initiative that all outputs are freely available to the community. It is also a basic aspiration, reflected in the *Digital Plan*, that Irish developments might support the efforts of other language communities, many of whom enjoy little or no State support or recognition. The kinds of challenges faced by minority and indigenous languages are often rather similar and the solutions for one language may be exactly right for many others. It is thus imperative that we learn from each other and that we develop mechanisms for cooperation, sharing our experience, expertise and, where possible, resources.

7. Acknowledgements

This work is supported by the Department of Culture, Heritage and the Gaeltacht with National Lottery funds. This forms part of the Government's 20-year Strategy for Irish 2010-2030.

8. References

ABAIR: the Irish language speech synthesizer - webpage with synthetic voices (2019). www.abair.ie

Barnes, E., Ní Chasaide, A. and Ní Chiaráin, N. (2018) The design and pre-testing of literacy and cognitive tasks in Irish and English, *Literacy Association of Ireland 42nd International Conference*, *Dublin, Ireland*, pp1 - 11

Dalton, M. and Ní Chasaide, A. (2005). Tonal alignment in Irish Dialects. *Language and Speech* 48, 441-464.

Dorn, A. and Ní Chasaide, A. (2015). Sentence mode differentiation in four Donegal Irish varieties. *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow, UK: the University of Glasgow. Paper number 0482, 5 pp. ISBN 978-0-85261-941-4. Retrieved from http://www.icphs2015.info/pdfs/Papers/ICPHS0482.pdf

Evans, G. (2018). Report on Language Equality in the Digital Age. Retrieved online 22/11/2019 from: http://www.europarl.europa.eu/doceo/document/A-8-2018-0228_EN.html

Mac Thomáis, U. (2018). "Tá guth-aithint an Bhéarla ag déanamh Béarlóirí de Ghaeilgeoirí, mé féin ina measc!" Article from tuairisc.ie. Retrieved from: https://tuairisc.ie/ta-guth-aithint-an-bhearla-ag-deanamh-bearloiri-de-ghaeilgeoiri-me-fein-ina-measc/

McGuirk, R. (2015). *Exploration of the Use of Irish Language Synthesis with a Screen Reader in the Teaching of Irish to Pupils with Vision Impairment*. Unpublished M.Phil. dissertation. Trinity College Dublin, Ireland.

Moseley, C. (Ed.). (2010). Atlas of the world's languages in danger (3rd ed.).Paris: UNESCO Publishing. Retrieved from: http://www.unesco.org/culture/en/endangeredlanguages/atlas

Murphy, A., Yanushevskaya, I., Ní Chasaide, A. and Gobl, C. (2019). The role of voice quality in the perception of prominence in synthetic speech, *INTERSPEECH 2019*, *Graz, Austria*, 2019, pp2543 - 2547

Ní Chasaide, A. (1999). Irish. *Handbook of the International Phonetic Association: a guide to the use of the international phonetic alphabet.* Cambridge: Cambridge University Press, 111-116.

Ní Chasaide, A., Ní Chiaráin, N., Berthelsen, H., Wendler, C., Murphy, A., Barnes, E. and Gobl, C. , (2019). Leveraging Phonetic and Speech Research for Irish Language Revitalisation and Maintenance, *ICPhS the 19th International Congress of Phonetic Sciences*, *Melbourne, Australia*, pp994 - 998

Ní Chiaráin, N. and Ní Chasaide, A. (2018). An Scéalaí: synthetic voices for autonomous learning. In: Taalas, P., Jalkanen, J., Bradley, L., Thouësny, S. (eds), *Future-proof CALL: language learning as exploration and encounters – short papers from EUROCALL 2018*, 230-235.

Ní Chiaráin, N. and Ní Chasaide, A. (2020). The Potential of Text-to-Speech Synthesis in Computer-Assisted Language Learning : A Minority Language Perspective In Andujar, A. (ed) *Recent Tools for Computer-and Mobile-Assisted Foreign Language Learning,* Hershey, PA, IGI Global, pp.149-169.

O'Reilly, M. and Ní Chasaide, A. (2016). Modelling the timing and scaling of nuclear pitch accents of Connaught and Ulster Irish with the Fujisaki model of intonation. *Proceedings of the 8th International Conference on Speech Prosody*, Boston, Massachusetts, pp. 355–359. DOI: 10.21437/SpeechProsody.2016-73

Yanushevskaya, I., Ní Chasaide, A. and Gobl, C. (2017) Cross-speaker variation in voice source correlates of focus and deaccentuation, *INTERSPEECH 2017*, *Stockholm, Sweden*, pp1034 - 1038

Proceedings, pages 182–185
Paris, UNESCO Headquarters, 5-6 December, 2019.

Spoken Language Technology for North-East Indian Languages

Viyazonuo Terhiija[1], Priyankoo Sarmah[1,2], Samudra Vijaya[2]
[1]Department of Humanities and Social Sciences
[2]Centre for Linguistic Science and Technology
Indian Institute of Technology Guwahati
Guwahati 781039, Assam, INDIA

{viyazonuo, priyankoo, samudravijaya}@iitg.ac.in

Abstract

The North-East India hosts a myriad of languages that belong to three different language families and that have salient phonological inventories. Officially, of the 99 non-scheduled languages, 60 are spoken in North-East India. Among these 60, 34 languages have more than 50,000 speakers. However, these languages do not have enough linguistic resources for language technology development. Many of them do not even have detailed linguistic descriptions that would be helpful in understanding the challenges that lay ahead in building language technology based tools in the languages. In this paper, we provide an overview of the languages of North-East India and outline a few phonetic features distinct to some of the languages. We also provide an overview of the attempts to build spoken language technology for some of these languages and finally we conclude by outlining the challenges in building speech technology in these languages and suggest some approaches to overcome the challenges.
Keywords: North-East India, Speech technology, Languages, Phonetic features

Cayie

North-East India nu die kekreikecii kekra baya. Die siiko die kikru se nu ba mu siikoe puo pfhephra kemeyie kekra baya. Diezho nu ba kemo die hiethepfiiethepfii donu die hiesorou-e North-East India nu puya mu siiko donu die seredia-e mia nyie hiepengou mese pu tuoya. Derei die hako chii kehielie kevi cha ba mo. Die hako puo dze puo se puo nyi di u khruohipie u mhodziitsatie u chiekezhiiko die chii kehielie kevi sikelie cha tuo mo. Leshii hau nu, North East India nu die pete donu rei die huo se par puo kepu pfhephra kekreikecii se kezashii. Siisie die-e kikemhie di chii kehie morosuo shi ikecii rei kezashii. Thelanu, die u chiekezhii mu kikemhie di u chiekezhii die siiko mho kuolietuo shi ikecii rei kezashii.

1. Introduction

The North-East India is part of India constituted of eight provinces with a population of 45 million. In spite of representing only 4% of the population of India, this area is linguistically diverse area with three major language families represented by native speakers of over 200 languages. While Indo-European and Austro-Asiatic languages are spoken in the area, there is a large number of Tibeto-Burman languages spoken in this area. There are only two major Indo-European languages, spoken as native languages in this area, namely Assamese and Bengali. However, the total number of speakers of these two languages is 27 million. In other words, the remaining 18 million speakers share the remaining 200 odd languages in the region. This makes the linguistic situation of the area extremely complicated, as several languages in the area are considered minority languages, and as a result of that there are not much linguistic resources available in these languages. The lack of such resources stand out as the first block in technology development in the languages of North-East India. Additionally, as the languages with smaller number of speakers belong to Tibeto-Burman or Austro-Asiatic languages, the linguistic features are quite distinct from other Indian languages. Hence, the approaches in developing speech technology in major Indian languages, such as Hindi or Bengali, may not be entirely applicable in the North-East Indian languages.

This paper is organized as follows. Section 2. provides an overview of the languages of North-East India. Section 3. provides examples of some of the phonetic features of North-East Indian languages that require special attention in building language technology, Section 4. provides a summary of the language technology developments in the North-East Indian languages and finally Section 5. discusses the way forward and concludes the paper.

2. Languages of North-East India

The eighth schedule of the constitution of India lists 22 languages as official languages, also called 'scheduled languages'. While English is not one of them, it is considered as a subsidiary official language (CoI, 1950). The constitution, however, does not prevent the states or provinces in India from choosing another languages, apart from the 'scheduled languages', as official language of the state. Apart from that the census of India has considered 99 languages as 'non-scheduled languages' that have more than 10,000 speakers each.

Table 1 shows the number of scheduled and non-scheduled languages spoken in North-East India, according to the census of India. The rightmost column lists the number of languages, subsumed under the scheduled and non-scheduled categories, that have more than 10,000 speakers. Of the total 90 mother tongues, 6 are Indo-European, 4 are Austrosiatic and 80 are Tibeto-Burman. As seen in the Table, the North-East Indian region has several Tibeto-Burman languages (Van Driem, 2018). Figure 1, shows the distribution of Tibeto-Burman languages in the world with North-East India hosting a sizeable number of them. The official languages of the eight states of North-East India are provided in Table 2 (Nag, 1963; Meg, 2005; Sik, 1977; Man, 1979;

Tri, 1964; Sik, 1977; Ass, 1960).

Table 1: Distribution of languages across North-East India

Languages	Number	Incl. mother tongues
Scheduled	4	10
Non-scheduled	60	80

Figure 1: Geographical distribution of the major Tibeto-Burman languages, argued to be Trans-Himalayan languages, (Van Driem, 2018). Each dot represents the putative historical geographical centre of each of 41 major linguistic subgroups. Source: (Van Driem, 2018)

Table 2: Official state languages across North-East India

States	Official languages
Arunachal Pradesh	English
Assam	Assamese, Bengali, Bodo and English
Meghalaya	English, Khasi and Garo
Manipur	Meiteilon, English
Nagaland	English
Tripura	Bengali, English and Kokborok
Mizoram	Mizo and English
Sikkim	Nepali, Bhutia, and English

3. Features of North-East Indian Languages

As seen in Table 1, the majority of languages spoken in the region are Tibeto-Burman languages. The Tibeto-Burman languages have their own distinct phonological features, some of which will be discussed in the sections below. Many of these features are not commonly found and hence, they emerge as linguistic challenges to deal with in speech technology development.

3.1. Lexical tones in North-East India

Tibeto-Burman languages are known to be tonal. There are only a few Tibeto-Burman languages that do not have lexical tones. As the majority of languages spoken in North-East India belong to the Tibeto-Burman family, almost all of them, barring a few, have lexical tones. Lexical tones are generally classified into two major groups namely, register tones and the contour tones (Yip, 2002). The range of tonal contrast varies in North-East India ranging from atonal to five or more and the inventory includes both register and contour tones. While, Bodo is reported to be a two-tone system (Sarmah, 2004), Mizo has four lexical tones in its inventory (Sarmah and Wiltshire, 2010). Acoustic studies have been conducted on Ao, Angami, Bodo, Dimasa, Mizo, Paite, Poula, Rabha, Tiwa etc. and the types and acoustic properties of lexical tones in these languages are fairly well understood. However, this still leaves out quite a number of tone languages in the region of which not much is known. In case of language technology development for tone languages, incorporation of tone information is of utmost importance as it improves recognition by disambiguating words. Several works have shown the advantage of incorporating tonal information in the development of Automatic Speech Recognition (ASR) systems in tone languages (Hu et al., 2014; Metze et al., 2013). Moreover, tone modeling needs to be exhaustive as it is also noticed that tones and segments interact with each other in a predictable manner (Lalhminghlui et al., 2019; Coupe, 1998; Sarmah, 2009).

3.2. Voiceless nasals

While nasals are known to be phonemically voiced, several Tibeto-Burman languages, such as Burmese, are known to phonemically contrast between voiced and voiceless nasals. Tibeto-Burman languages spoken in North-East India, namely, Mizo and Angami, also show evidence for voicing distinction in nasals. Mizo voiceless nasals are primarily voiceless with a bit of voicing towards the end of the nasal segment. On the other hand, Angami voiceless nasals are entirely voiceless with aspiration at the end (Bhaskararao and Ladefoged, 1991). Hence, Angami contrasts between /m/ and /m̊^{h}/, /n/ and /n̊^{h}/,/ŋ/ and /˚ŋ^{h}/. In Mizo, the contrastive nasals are: /m/ and /m̊/, /n/ and /n̊/,/ŋ/ and /˚ŋ/. Such phonetically similar segments my pose challenges in speech technology development.

3.3. Fricative Aspiration

Another recently reported phenomenon in Tibeto-Burman languages in North-East India is the aspiration of the fricative sounds. Bodo and Rabha have reported the existence of aspiration associated with voiceless, alveolar fricatives, /s/ (Sarmah and Mazumdar, 2015; Rabha et al., 2019). It is reported that phone recognition in Rabha is better when aspiration in fricatives is taken into account using acoustic features such as strength of excitation (SoE) and variance of successive epoch intervals (VSEI) (Rabha et al., 2019).

4. Spoken Language Technologies

A brief summary of the spoken language technologies developed for the languages of North-East India is presented in this section. The 3 'scheduled' languages of North-East India (Assamese, Bodo, Manipuri) can be considered as under-resourced languages. Most of the speech technologies were developed for these languages, thanks to the support from the Government of India. Most other languages

Table 3: Speech technologies developed for the languages of North-East India

Language	**ISO** 639-3 **code**	**No. of speakers**	**Speech Systems**
Angami	njm	152,796	ASR
Ao	njo	260,008	DID
Assamese	asm	15,311,351	ASR,TTS, PE
Bodo	brx	482,929	ASR,TTS
Khasi	kha	1,431,344	DID
Manipuri	mni	1,761,079	ASR,TTS, KWS
Mizo	lus	830,846	ASR,PE
Sora	srb	5,900	ASR

of North-East India can be considered as zero-resource languages. The spoken language technologies developed for the languages of North East India are listed in Table 3. The ISO 639-3 code as well as the number of persons speaking the language (cen, 2011) are also given in the Table. A brief account of various speech systems implemented for these languages is given below.

4.1. Angami

Angami (also Tenyidie) language is spoken in the state of Nagaland. A preliminary Automatic Speech Recognition (ASR) system was developed using speech data of sentences read by 11 native speakers (Terhiija et al., 2019). The Word Error Rate (WER) of the ASR system on the training data was under 5%. In a 'leave one speaker out' cross validation experiment, the average WER of the ASR system using context independent phone Hidden Markov Model (HMM) was 17.3%.

4.2. Ao

Ao is spoken in Nagaland. A Gaussian Mixture Model (GMM) based Dialect Identification (DID) system was implemented to identify two Ao dialects, namely, Changki and Mongsen (Tzudir et al., 2018). Augmentation of spectral features with tonal features resulted in better DID accuracy.

4.3. Assamese

Several speech systems were implemented for Assamese, a 'scheduled' language, spoken in the state of Assam. An Assamese ASR system was implemented using speech speech data from 209 speakers. The ASR system that employed a Deep Neural Network (DNN) along with HMM yielded with the lowest WER of 12.4% (Deka et al., 2019b). A Phonetic Engine (PE) was implemented with a phone recognition accuracy of 47.31%, 45.30%, 36.13%in reading, lecture and conversation modes respectively (Sarma et al., 2013). A text-to-speech (TTS) system using a DNN was developed. The quality of the synthesized speech was distinctively better than that of the speech synthesized by the GMM-HMM based TTS system (Deka et al., 2019a).

4.4. Bodo

Bodo, a 'scheduled' language, is spoken in the state of Assam. A GMM-HMM based ASR system was implemented in 2014 (Laba Kr. Thakuria, 2014). A TTS system for Bodo was built using the concatenative approach (IITG TTS group, 2013).

4.5. Khasi

Khasi is a language spoken in the state of Meghalaya. A GMM based Dialect Identification system was implemented that recognises the dialect of the input Khasi speech as one of the two dialects: Khynriem or Bhoi-Jirang with an accuracy of 97% (Arjunasor Syiem, 2016).

4.6. Manipuri

Manipuri (also Meiteilon) is spoken in the state of Manipur. Development of speech technology for Manipuri language in form of ASR and Keyword Search (KWS) system is reported. The authors collected and transcribed telephonic read speech data of A speech database of over 90 hours telephonic speech from more than 300 speakers was created for implementation of an KeyWord Spotting (KWS) system as well an ASR system (Patel et al., 2018b). The WER of the DNN-HMM based ASR system was 13.5%. The equal error rate of the KWS system was 7.64% (Patel et al., 2018a). A HMM based TTS system for Manipuri language was implemented (IITG speech group, 2013). A toolkit to build TTS enabled one to build TTS system in many languages including Manipuri (Ghone et al., 2017).

4.7. Mizo

Mizo is spoken in Mizoram state. Phonetic Engine (Dey et al., 2017) as well as ASR systems (Kothapalli et al.,), (Dey et al., 2018) were implemented for Mizo language. The phone recognition rate of the PE Mizo phonetic engine was 13.9% when DNN-HMM acoustic model was used in conjunction with language model (Dey et al., 2017). The WER of DNN-HMM based ASR system for clean speech was 13% (Dey et al., 2018).

4.8. Sora

Sora is mainly spoken in the states of Orissa and Andhra Pradesh. In the state of Assam, there are 5,900 Sora speakers whose ancestors migrated to Assam in the 19th century. A DNN-HMM based ASR system was implemented, and had a WER of 13.9% (Chakraborty et al., 2018).

5. Conclusion

This paper presented salient features of languages spoken in North-East India, and gave an account of speech systems developed for some of these languages. The zero-resource status of the most of these languages is a major barrier in enabling people of all strata to reap the benefit of language technology. A good study of the linguistic properties of these languages would set a foundation for building spoken language systems for these languages via transfer of knowledge from related languages.

References

Arjunasor Syiem, Gaurab Krishnan Deka, T. I. L. J. S. (2016). Khasi dialects identification based on gaussian mixture model. *Int. J. Engg. Sci. Computing*, 6(4):3882–3885.

(1960). *The Assam Official Language Act, 1960*.

Bhaskararao, P. and Ladefoged, P. (1991). Two types of voiceless nasals. *Journal of the International Phonetic Association*, 21(2):80–88.

(2011). *Office of the Registrar General & Census Commissioner India, 'Statement-1 Part-B Languages not specified in the eighth schedule (non-scheduled languages)*.

Chakraborty, K., Horo, L., and Sarmah, P. (2018). Building an automatic speech recognition system in sora language using data collected for acoustic phonetic studies. In *SLTU*, pages 239–242.

(1950). The Constitution of India: Part XVI: Official language.

Coupe, A. R. (1998). The acoustic and perceptual features of tone in the tibeto-burman language ao naga. In *Fifth International Conference on Spoken Language Processing*.

Deka, A., Sarmah, P., Samudravijaya, K., and Prasanna, S. (2019a). Development of assamese text-to-speech system using deep neural network. In *2019 National Conference on Communications (NCC)*, pages 1–5. IEEE.

Deka, B., Sarmah, P., and Vijaya, S. (2019b). Assamese database and speech recognition. *22nd Oriental-COCOSDA, Cebu, Philippines*.

Dey, A., Lalhminghlui, W., Sarmah, P., Samudravijaya, K., Prasarma, S. M., Sinha, R., and Nirrnala, S. (2017). Mizo phone recognition system. In *2017 14th IEEE India Council International Conference (INDICON)*, pages 1–5. IEEE.

Dey, A., Sarma, B. D., Lalhminghlui, W., Ngente, L., Gogoi, P., Sarmah, P., Prasanna, S. R. M., Sinha, R., and S.R., N. (2018). Robust mizo continuous speech recognition. In *Proc. Interspeech 2018*, pages 1036–1040.

Ghone, A., Nerpagar, R., Kumar, P., Baby, A., Shanmugam, A., Mukundan, S., and Murthy, H. (2017). Tbt(toolkit to build tts): A high performance framework to build multiple language hts voice. 08.

Hu, W., Qian, Y., and Soong, F. K. (2014). A dnn-based acoustic modeling of tonal language and its application to mandarin pronunciation training. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3206–3210, May.

IITG speech group, (2013). *IIT Guwahati Text-to-Speech Synthesis for Manipuri Language*.

IITG TTS group. (2013). Text-to-speech synthesis for bodo language. `http://www.iitg.ac.in/cseweb/tts/tts/Bodo/`.

Kothapalli, V., Sarma, B. D., Dey, A., Gogoi, P., Lalhminghlui, W., Sarmah, P., Prasanna, S. M., Nirmala, S., and Sinha, R.). Robust recognition of tone specified mizo digits using cnn-lstm and nonlinear spectral resolution.

Laba Kr. Thakuria, Purnendu Acharjee, A. D. P. T. (2014). Bodo speech recognition based on hidden markov model toolkit(htk). *Int. J. Sci. Engg. Res.*, 5(1):339–343.

Lalhminghlui, W., Terhiija, V., and Sarmah, P. (2019). Vowel-tone interaction in two tibeto-burman languages. *Proc. Interspeech 2019*, pages 3970–3974.

(1979). *The Manipuri Official Language Act, 1979*.

(2005). *The Meghalaya Official Language Act, 2005*.

Metze, F., Sheikh, Z. A., Waibel, A., Gehring, J., Kilgour, K., Nguyen, Q. B., and Nguyen, V. H. (2013). Models of tone for tonal and non-tonal languages. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 261–266. IEEE.

(1963). *The Nagaland Official Language Act, 1963*.

Patel, T., Krishna, D., Fathima, N., Shah, N., Mahima, C., Kumar, D., and Iyengar, A. (2018a). An automatic speech transcription system for manipuri language. In *Interspeech*, pages 2388–2389.

Patel, T., Krishna, D., Fathima, N., Shah, N., Mahima, C., Kumar, D., and Iyengar, A. (2018b). Development of large vocabulary speech recognition system with keyword search for manipuri. In *Interspeech*, pages 1031–1035.

Rabha, S., Sarmah, P., and Prasanna, S. M. (2019). Aspiration in fricative and nasal consonants: Properties and detection. *The Journal of the Acoustical Society of America*, 146(1):614–625.

Sarma, B. D., Sarma, M., Sarma, M., and Prasanna, S. M. (2013). Development of assamese phonetic engine: Some issues. In *2013 Annual IEEE India Conference (INDICON)*, pages 1–6. IEEE.

Sarmah, P. and Mazumdar, P. (2015). Aspiration in alveolar fricatives in bodo. In *ICPhS*.

Sarmah, P. and Wiltshire, C. R. (2010). A preliminary acoustic study of mizo vowels and tones. *J. Acoust. Soc. Ind*, 37(3):121–129.

Sarmah, P. (2004). *Some aspects of the tonal phonology of Bodo*. Ph.D. thesis, English and Foreign Languages University, Hyderabad.

Sarmah, P. (2009). *Tone systems of Dimasa and Rabha: a phonetic and phonological study*. University of Florida.

(1977). *The Sikkim Official Language Act, 1977*.

Terhiija, V., Sarmah, P., and Vijaya, S. (2019). Development of speech corpus and automatic speech recognition of angami. *22nd Oriental-COCOSDA, Cebu, Philippines*.

(1964). *The Tripura Official Language Act, 1964*.

Tzudir, M., Sarmah, P., and Prasanna, S. M. (2018). Dialect identification using tonal and spectral features in two dialects of ao. In *SLTU*, pages 137–141.

Van Driem, G. (2018). The East Asian linguistic phylum: A reconstruction based on language and genes. *The Asiatic Society*, 60(4):1.

Yip, M. (2002). *Tone*. Cambridge University Press.

Proceedings, pages 186–189
Paris, UNESCO Headquarters, 5-6 December, 2019.

Nenek: Digital Self-documentation for Minority and Under-resourced Languages

Anuschka van 't Hooft, José Luis González
Universidad Autónoma de San Luis Potosí, Cinvestav-Tamaulipas
Av. Industrias 101-A, Fracc. Talleres, 78399 San Luis Potosí, S.L.P., Mexico;
Km. 5.5 carretera Cd. Victoria-Soto la Marina, 87130 Cd. Victoria, Tamps, Mexico
avanthooft@uaslp.mx, jgonzalez@tamps.cinvestav.mx

Abstract

The Nenek platform was designed to support speech communities in their language documentation efforts. Its features and linguistic tools enable speakers to work together in virtual communities and create, document, archive, and mobilize language resources. These, in turn, can be used to manufacture more complex multimedia language resources, all of which are available under the "open archive" principle. The four-stage cyclical management model and the use of virtual communities throughout the documentation process facilitate the self-documentation of language and culture and provide a space for speakers to discuss about and in their language. Also, it makes the minority language visible on the Internet.

Keywords: self-documentation, language resources management, virtual communities

Résumé

An ts'ejkadh mulkux kaw Nenek k'wajat t'ajadh abal ka tolmiyat an kwenchal tin tsáb abal kin jila' dhuchadh in kawintal. In walkixtal ani in eyextalabh tin kawintal in t'ajal ka t'ojon ti jun yanel an kwenchalchik axi k'wajat ti al an tsu'udh kaw abal kin tsalpay, dhucha', dhaya' ani ka wat'baxin patal in kawintal. Patal axe' xi tolmixtalab k'wajat japidh abal jita' kin le'na' ba jun i tsakam muke'. Axe' xi tsalpaxtalab abal kin mulkuw an tolmixtalab in t'ajal ádhik an kwetem-dhuchnel kawintalab ani an biyal t'ajnel, in t'ajal jun i tamkuntalab abal ka t'ilmaxin an ebchalabchik tin kwetem kawintal, ani abal ka tejwamej axe' xi kawintalab ti al an buk'ux kaw.

1. Technological support for language documentation and revitalization efforts

Language documentation (Himmelmann, 2006) is one of the main answers to language endangerment, as it aims to compile and preserve linguistic primary data of communicative events and creates interfaces to bring about the study of these data. One of the two leading paradigms of documentary practices is active documentation or documentation oriented to the community (Flores Farfán and Ramallo, 2010). Active documentation is often community-based and linked to efforts to maintain and revitalize endangered languages.

We developed an online collaborative strategy for speech communities to engage in self-documentation projects (Quatra, 2011) of their language and culture. In these projects, speakers of minority languages and under-resourced languages are the main investigators, compilers and users of language resources. These language resources are either audios, texts, videos or images, and can be combined and worked on to create multimedia materials.

Our project is called Nenek, which is also the name of the online platform that operates as a virtual community and was designed to support the self-documentation activities of the Huastec[1] speech community (van 't Hooft and González, 2014). In Huastec, "nenek" is an informal greeting, which aims to be inviting, plus it provides the platform with a recognizable local identity that make speakers feel part of a community. In this paper, we present the features of the Nenek platform (Nenek, 2019) and discuss its qualities to support self-documentation for minority language speakers.

2. The Nenek Platform: Language Resources Management Model

Nenek supports the creation of virtual communities of minority language speakers on the Internet. This platform includes a set of tools that enables users to work collaboratively on language documentation tasks, build lexicographic assets and produce new language resources (van 't Hooft and González, 2014; González, *et.al*, 2017a, 2017b). Here, it shifts from the common five-stage process of language documentation -recording, capture, analysis, archiving and mobilization- (Austin, 2006) to a four-stage cyclic management model to control the acquisition of existing materials in the target language and the manufacturing and archiving of new language resources, as well as their distribution within the virtual community and to the general public.

(1) In the acquisition stage, already existing language materials are either automatically extracted from the web by a crawler or received through donations from users who participate in a monolingual virtual community.

(2) In the manufacturing stage (merging recording, capture, and analysis), the speakers collaboratively document these acquired language materials, creating metadata and annotations. They also manufacture and document new language resources, either exclusively with their own means (e.g. filming a ceremony or recording a story) or using the acquired language resources (e.g. combining grandmother's knowledge on local gastronomy with photos from the virtual community to create a recipe book). All resources are discussed and validated by the virtual community before entering the repositories and corpora.

[1] Huastec is a Mayan language spoken across communities and towns in the subtropical Gulf region of Mexico. This language, known by its speakers as Tének, Teenek or Tenec (due to its three writing systems), has three linguistic varieties (INALI, 2008) and at least 170.000 speakers (INPI, 2019).

(3) Meanwhile, Nenek creates repositories and computerized corpora of the language, which are exploited to generate linguistic tools such as spell checkers and e-dictionaries. Likewise, these tools enable the building of new language resources.
(4) The acquired and manufactured resources are published in the mobilization stage, either within the virtual community or publicly. All these language resources (in formats such as audios, videos, texts, photos and multimedia) are available on the Nenek platform as an e-library and can re-enter the manufacturing stage when used again to create more complex language resources.
A life cycle mapping scheme registers the transformations of the language resources at each of the stages of the language resources management cycle. This scheme also traces the utilization and diffusion of each resource that is produced by the virtual community.

3. The Nenek Platform: Virtual Communities

Speakers of minority languages are increasingly exposed to forced migrations, which are often the result of economic pressures and inequalities, but also of environmental disasters and social and political conflicts. These migration processes contribute to language loss. Our proposal to work through virtual communities brings these migrants -mostly young people or heritage speakers- into the documentation project, giving them spaces to use their language as well as a task in the conservation of their language and culture.
In a virtual community (VC), the participants are involved in social or emotional interpersonal interactions and have access to information. What makes users gather and operate as a community is both the type of relations they establish as well as the characteristics of that network. In a VC, users are interdependent: resources are transferred from the links existing among members and are based on their common interests (Gupta and Kim, 2004).
We identified the most effective strategies to encourage speakers' participation in the native language on the Internet (González Compeán *et.al*, 2017a). After creating pages on the most popular social networks (Facebook, Twitter), the life cycle of the VC was steered and monitored through the postings of three types of publications (announcements, debates, and pep talks) with a clear graphical identity, in a specific order and with regular time intervals. At the same time, speakers were invited to join the Nenek platform, which is also a VC, and collaborate in the acquisition, manufacturing and mobilization stages of the documentation process.
On the Nenek platform, the VC includes functionalities such as profiles, work groups and contents management, as well as tools that allow the speakers to create web pages and blogs in which they can contribute to the repository building by sharing and discussing texts, audios, images and videos. The documentation activities are carried out collaboratively and in a cyclic process that starts when the speakers propose a task for a work group and store their materials in one of the repositories. Then, either the speakers' communication or input of materials returns to the VC after a categorization and consensus polling procedure (validation process) carried out by the collaborators.

4. Discussion

Nenek was developed to support speakers' engagement in documentation activities while discussing among themselves about their language. It is a unique design that can be replicated in other documentation efforts of minority languages, since it can be localized and installed for minority languages in a relatively short time (Manuals are available online to facilitate its use). Among the special features of our project are its focus on active documentation, its cyclical resources management, and the use of virtual communities throughout the documentation process.

4.1 Active documentation

As an active documentation project, Nenek was set off by members of the Huastec speech community, who designed the platform and started the documentation activities. The platform can only operate in a close collaboration with the speech community, including their wishes and needs in each project, and involving them directly in all stages of the documentation process. Nenek addresses differentiation and heterogeneity in the speech community (Grinevald and Bert, 2011), provides training for collaborators, and guides the documentation activities aiming for speakers to become custodians of their linguistic and cultural heritage (Czaykowska-Higgins, 2009; Furbee, 2010). Thus, Nenek creates language resources that are esteemed to be necessary or valuable by the speech community and also agree with academic standards of language documentation, which enables their use by the scientific community.
The Nenek platform is flexible and can be adapted to the needs of speakers and their language. Not all features have to be used, and Nenek can operate next to other documentation or revitalization tools and practices. Also, it is versatile, in that it focuses on the documentation of language and culture and not only on linguistic documentation, as speakers often see language as being intricately interwoven with their cultural heritage (Franchetto, 2006). Accordingly, it supports the documentation of materials that deal with more cultural aspects of the speech community, such as dances and music, in formats that do not necessarily contain speech, such as photos.

4.2 Resources management

Nenek's resources management scheme incorporates an acquisition stage to collect and digitize donations and already existing materials in the target language on the Internet. For under-resourced languages, these materials are often dispersed and difficult to find (González Compeán *et.al*, 2017b). The inclusion of this stage in the documentation process provides a first e-library with language resources, which may also contain sources with information about the language and culture, as well as some linguistic tools (e-dictionaries, spell checkers), and is vital to attract speakers to the documentation project and initiate the virtual community (Garber, 2004; Iriberri and Leroy, 2009).

The management scheme is cyclical, in that manufactured language resources that are validated by the speakers are automatically mobilized and can be used again to create more complex ones. This mobilization through the VC enhances the use of language resources, fuels discussions among speakers about the contents and about their language, and strengthens the VC.

Our resources management scheme allows the building of linguistic tools that support speakers in their documentation activities, as they require a degree of literacy (for annotations, metadata, messages to the CV). More often than not, minority language speakers are short of experience with reading and writing in their mother tongue in any medium and they develop literacies during the project, which must be carefully discussed (Lüpke, 2011).

4.3 Virtual communities

Nenek uses virtual communities in all stages of the project, which is challenging but feasible. Young speakers of minority languages are present on the Internet and are participating in social networks. Some of them already are involved in digital language activism projects (Llanes-Ortiz, 2016) to promote their languages. Migrant speakers are especially keen to collaborate, since it connects them with people who share the same cultural background. When creating new language resources, their activities usually engage older members of their families and home villages, so that the whole speech community becomes involved and a broad range of verbal expressions can be addressed.

The key factors in the success of a VC are the constant generation of contents and the availability of those contents online. Our strategy to improve the life cycle of the VC requires a continuous mobilization of the acquired and manufactured language resources (González Compeán *et.al*, 2017a). In order to be successful, the contents on the platform, in the format of multimedia materials, should address, in the first place, community needs and interests in the language and culture.

With the use of VCs as a means to develop language documentation projects, language resources are acquired, manufactured, mobilized and discussed on the Internet, which brings about a greater visibility of these minority languages. For young people, it can be stimulating to see that their language is consistent with a modern global network society like the Internet and that the minority language can be used to express oneself in this new system of communication. Therefore, it also enhances processes of digital self-determination (McMahon, 2013).

5. Ethical issues

Active documentation recognizes that the situation and position of the world's languages are an expression of historical processes in which some languages have had economic, cultural and political advantages over others. Different from the type of documentation in which solely scientific questions are addressed, it conceives of documentation as a means to assist speech communities in their efforts to maintain and revitalize their language (Flores Farfán and Ramallo, 2010), and as a way to help guarantee the use of the mother tongue as a fundamental human right (Unesco, 2003). Here, several ethical issues have to be addressed, of which we can only mention a few.

We aspire to improve the weak situation and position of minority languages on the Internet and address the linguistic rights of people to participate on the Internet on their own terms, thus contributing to revitalization processes. For one, the Nenek platform is localized in the minority language which is the focus of the project. Also, the members of the virtual community communicate in their mother tongue while collaborating online and not in the majority language. This monolingual immersion is a logical consequence of our aim to create more horizontal relations among all participants in the project, at the same time expanding the domains of the use of the minority language and creating more visibility of this language on the Internet.

Additionally, we recognize the demand of speakers to have access to the outcome of the documentation project. Nenek is an open archive with multiple repositories, which are available to speakers. Automatic URLs guarantee the rights of the authors and participants of each language resource.

Furthermore, Nenek recognizes the oral expression as an essential component in the transmission of linguistic and cultural knowledge and as a fundamental element to express and protect the linguistic and cultural diversity (Maffi, 2003), including all varieties of the language, and favoring the compilation and mobilization of oral language resources. The written expression is used in new media, social networks and other internet-based resources, accepting all existing written norms for languages that are not standardized. Standardization of the language is not held to be a prerequisite to promote literacy processes.

This way, the Internet can become a strong ally in preserving and revitalizing minority languages, as speakers find here linguistic situations and materials that allow and even vindicate the use of the language, and they widen the spaces for expression in their mother tongues. However, in order to be effective, these virtual initiatives should be accompanied by other efforts to preserve the languages (according to the sociolinguistic diagnosis of each speech community), as well as by appropriate linguistic policies to ensure and promote their autonomous development.

6. Conclusions

Nenek's technological support fosters the empowerment of indigenous peoples in taking care of their linguistic and cultural heritage, making it a project for and by native speakers. At present, more than 3,300 Huastec speakers are actively involved in the Nenek project and decide together on the activities they want to develop to document and promote their language. Their Internet activity generates materials in the Huastec language and enables us to retrieve and document different types of sources.

In Nenek, the complete resource life cycle happens on the Internet, which represents a source to get new resources (acquisition stage), a space to produce (manufacturing stage) and store (archiving stage) new resources and a site to publish acquired and manufactured language resources (mobilization stage).

To the members, contributing in a VC through collaborative action in the production of digital contents is an opportunity to articulate, transmit and discuss knowledge and traditional practices. By placing their language on the Internet, speakers participate to expand

their linguistic heritage into new domains of usage and build a forum to discuss issues about their language and culture. This participation makes the language and culture visible in a virtual medium like the Internet. Nenek is an example of the important role the Internet plays in the current repositioning of indigenous and minority languages as opposed to the dominant languages.

7. Acknowledgements

The Nenek project was sponsored by CONACYT, the Mexican Council for Science and Technology, through grant CB-2012-180863. A technology transfer to INALI, the National Institute of Indigenous Languages, makes it possible to support speech communities in Mexico who want to use our platform and start their own language documentation project. We thank Alejandra Santiago Bautista for translating the abstract of this paper into the Huastec language.

8. References

Austin, P. K. (2006). Data and Language Documentation. In J. Gippert, N. P. Himmelmann and U. Mosel (eds.), Essentials of Language Documentation, pages 87-112. Berlin / New York: Mouton de Gruyter.

Czaykowska-Higgins, E. (2009). Research Models, Community Engagement, and Linguistic Fieldwork: Reflections on Working with Canadian Indigenous Communities. *Language Documentation and Conservation* 3(1):15-50.

Flores Farfán, J. A. and Ramallo, F. (2010). Exploring Links Between Documentation, Sociolinguistics and Language Revitalization: An Introduction. In J. A. Flores Farfán and F. Ramallo (eds.), New Perspectives on Endangered Languages. Bridging Gaps Between Sociolinguistics, Documentation, and Language Revitalization, pages 1-12. Amsterdam / Philadelphia: John Benjamins Publishing Company.

Franchetto, B. (2006). Ethnograohy in language documentation. In J. Gippert, N. P. Himmelmann and U. Mosel (eds.), Essentials of Language Documentation, pages 183-211. Berlin / New York: Mouton de Gruyter.

Furbee, N. L. (2010). Language Documentation. Theory and Practice. In L. A. Grenoble and N. L. Furbee (eds.), Language Documentation. Practices and Values, pages 3-24. Amsterdam / Philadelphia: John Benjamins Publishing Company.

Garber, D. (2004). Growing Virtual Communities, *International Review of Research in Open and Distance Learning* 5(2):1-7.

González Compeán, J. L., van 't Hooft, A., Carretero Pérez, J., and Flores Martínez, L. (2017a). La introducción de la lengua huasteca a internet. Una estrategia para crear comunidades virtuales en lenguas amerindias. *Comunicación y Sociedad* 28:131-153. http://www.comunicacionysociedad.cucsh.udg.mx/index.php/comsoc/article/view/6399 (Accessed: 21 November 2019).

González, J. L., van 't Hooft, A., Carretero Pérez, J. and Sosa Sosa, V.J. (2017b). Nenek: a cloud-based collaboration platform for the management of Amerindian language resources. *Language Resources and Evaluation*, 51(4):897-925. doi: 10.1007/s10579-016-9361-8

Grinevald, C. and Bert, M. (2011). Speakers and Communities. In P. Austin and J. Sallabank (eds.), The Cambridge Handbook of Endangered Languages, pages 45-65. New York: Cambridge University Press.

Gupta, S. & Kim, H. W. (2004). Virtual community: Concepts, implications, and future research directions. Proceedings of the Tenth Americas Conference on Information Systems, pages 2679-2687. New York.

Himmelmann, N. P. (2006). Language Documentation: What is it and What is it Good for? In J. Gippert, N. P. Himmelmann and U. Mosel (eds.), Essentials of Language Documentation, pages 1-30, Berlin / New York: Mouton de Gruyter.

Hooft, A. van 't and González Compeán, J. L. (2014). Collaborative Language Documentation: The Construction of the Huastec Corpus. In L. Pretorius, C. Soria and P. Baroni (eds.), Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014), pages 67-70. Reykjavik, Iceland, European Language Resources Association (ELRA).

INALI (2008). *Catálogo de las Lenguas Indígenas Nacionales: Variantes Lingüísticas de México con sus autodenominaciones y referencias geoestadísticas*. Mexico, Instituto Nacional de Lenguas Indígenas.

INPI (2019). *Atlas de los pueblos indígenas de México*. http://atlas.cdi.gob.mx/ Mexico, Instituto Nacional de Pueblos Indígenas (Accessed: 7 October 2019)

Iriberri, A. and Leroy, G. (2009). A Life-Cycle Perspective on Online Community Success. *ACM Comput. Surv.* 41(2), article 11. https://dl.acm.org/citation.cfm?id=1459356 (Accessed: 21 November 2019)

Llanes-Ortiz, G. (2016). Primeros pasos de estudio participativo: lenguas indígenas y medios digitales. https://rising.globalvoices.org/lenguas/2016/01/12/estudio-addli-primeros-pasos/ (Accessed: 21 November 2019)

Lüpke, F. (2011). Orthography development. In P. K. Austin and J. Sallabank (eds.) *The Cambridge handbook of endangered languages*, pages 312-336. New York: Cambridge University Press.

Maffi, L. (2003). The 'Business' of Language Endangerment: Saving Languages or Helping People Keep Them Alive. In H. Tonkin and T. Reagan (eds.), Language in the 21st Century, pages 67-86, Amsterdam / Philadelphia: John Benjamins Publishing Company.

McMahon, R. (2013). *Digital Self-determination: Aboriginal Peoples and the Network Society in Canada*. Ph.D.Thesis. Vancouver, Simon Fraser University.

Nenek (2019). Nenek. Platform for collaborative self-documentation of language and culture. http://nenek.inali.gob.mx (Accessed: 21 November 2019)

Quatra, M. M. (2011). 'Auto-documentación Lingüística': La Experiencia de una Comunidad Jodï en la Guayana Venezolana. *Language Documentation & Conservation,* 5:134-156.

UNESCO (2003). *Language Vitality and Endangerment*. UNESCO Ad Hoc Expert Group on Endangered Languages, 23 pp.

Proceedings, pages 190–192
Paris, UNESCO Headquarters, 5-6 December, 2019.

Indigenous Language Revitalization: The Context of Inpui Naga in Northeast India

Rajiandai Bariam

Naga People's Movement for Human Rights

217, Hostel Annexe, University of Hyderabad, Gachibowli, Hyderabad, Telangana. 500046

ajeeanbung@gmail.com

Abstract

In the past few decades, there had been an increasing trend of indigenous language revitalization around the world. But the dilemma of weighing the pros and cons in the process of revitalization invites critical engagement. This article presents some of the issues of language revitalization and also placed it in the context of the Inpui speakers of the Nagas from northeast India. The article conveyed that inspite of various challenges that confronted it, indigenous language revitalization and bilingual form of communication needed to be arranged through civil societies, state or international mechanisms.

Keywords: language shift, revitalization, bilingual,

Résumé

Sedamnu sangwan somni somthum lakpiya, baram baram ruan chong piringthoulatlang karam rakan kanu muwe. Dida heiba chuanna kakhanchuilang bathang amme. Hei om chuihiak heina Inpui tinu Naga ruan India om nisaksuakpekla lungnu ruan lambi thunknu balang kasem hiakthang nuwe. Heiba chuihiak na rinnuba, chong piringkalatnu heiya kungtaknu inja kadou peida, ani chonga, ruan lianthounu chongle bachu bachamlang, makebe kabia lang, meebing, sasaan khatle taleipuba ruanle kunga kut kaikarunga chuan tolang bathang amme diga rinnu we.

1. Introduction

1.1 Views on Indigenous language Revitalisation

The issue of indigenous language revitalisation had been a meeting point for scholarship and activism. Post world war II, it had occupied a crucial place among the indigenous people. The question of indigenous language literacy and documentation faced various contestations. Russell Means, the American Indian writer leader sets it clear in his speech about the tribal's uncomfortability with writing,

> "The only possible opening for a statement of this kind is that I detest writing. The process itself epitomizes the European concept of "legitimate" thinking; what is written has an importance that is denied the spoken. My culture, the Lakota culture, has an oral tradition, so I ordinarily reject writing. (Padel, 1996)

The discussion also revolved around "autonomous" versus the "ideological" models of literacy. (Street 1984) While the "autonomous" model considered indigenous language literacy to be "neutral" technologies that can be easily detached, the "ideological" model "concentrates on the social practice of reading and writing"[1]. According to Brian Bielenberg," literary discussions in the context of language revitalization should be looked at from an ideological model of literacy" (1999). Various texts showed the struggle and the challenges involved in the attempt and the process of revitalization of the indigenous language. In the context of the Navajo speakers, many felt that documentation could lead institutions to become more indigenous. But others argued that English literacy is the way to "get ahead" and indigenous language education could become an obstacle in the way of their children's future. (McCaulin, 1992; Bielenberg, 1999) But even after its indigenous language incorporation, the use of Navajo was limited to the schools and church which were the two powerful institutions that teaches and communicates the language in a way different from the traditional Navajo.

These challenges mentioned above showed the complexities that confronted the task of indigenous language revitalization. Bielenberg aptly reminded the stakeholders to be aware of the cultural repercussions that come with such task of revitalization. On the other hand, McLaughlin strongly voiced about empowerment through literacy. He also suggested three points to be reminded in undertaking such task, two of which advocated for long term commitment to hiring local individuals for teaching and an undoing of the status quo and prestige hierarchy of oral and written English (McLaughlin, 1995). Nancy Hornberger assured the beneficial use of a bilingual education stating that the use of indigenous language alongside dominant language results to dialogism, meaning-making and access to wider discourses (2006). According to Teresa McCarythy revitalizing indigenous language is a means of resistance in the face of homogenizing and standardizing force of globalization. (2003)

2. The Constitution of Life Through Language

Language is the means through which the indigenous people expressed themselves and chart out their way of life. In the context of the Nagas of Northeast India, the folktales surrounding Hornbill were a very significant way of shaping the tribal society towards kindness, and rooting a sense of justice in them. The folklores, and cultural events and traditions contain paradigms of social living.

[1] Bielenberg, 1999

But the process of revitalization had been scarce and inadequate or even worse at least among the Inpui.

3. Situation of the local language in formal Institutions

I attended a school where English was compulsory not only as the medium of teaching, but also communication among students and teachers. Local dialects were forbidden and repressed. The ability to communicate in English became a symbol of class and a form of status quo. A binary lens exposed the repression of indigenous dialects on one hand, and the popularization of the mainstream English language on the other. It is often propagated through a well-placed system, institutionalized and that eventually became a social norm.

4. Position of Inpui language and challenges

I belong to a tribe called Inpui from the state of Manipur in India. Inpui falls under the "unsafe" under the Intergenerational Language transmission and more so as it is a small speech community. It is also an undocumented language. (Based on the document submitted to the International Expert Meeting on UNESCO Programme Safeguarding of Endangered Languages Paris 10-12 March 2003)[2] With just around twelve thousand native speakers, the tribe is a minority among the Nagas, and they rapidly face threats not only from the institutionalized languages, but also from the majority tribes in close proximity to them. It has been told that many of the Inpui villages were assimilated into other larger tribes (Khumba 2012). And in the present, the rapid change in their cultural norms, social practices and ubiquity of the western culture and its dominant English language through the internet spaces, social space and institutionalized spaces had radical impact on the indigenous language and its articulation. With it also came the loss of their culture, folktales, social practices and meetings that were once spaces within which a language is learned and articulated.

In other words, with the loss of the indigenous spaces invaded by the new western cultural and language norms also brings the phenomenon of "language shift" or the loss of the indigenous language. The relationship between indigenous language and English has always been biased. In a post-colonial context, the odds are always against the indigenous languages. Without a systematic effort to protect them, the people to whom they are important could be stripped of all their indigeneity.

The present education system hasn't been inclusive with the indigenous people nor about their knowledge systems. They are often kept out of the ambit of the present forms of what constitutes knowledge. When indigenous knowledge and practices are considered as inferior or unworthy of being considered knowledge, where should the countervailing mechanism rest? It begins with documenting and coding their language and their knowledge. State and International institutions need to make space for these languages and its culture that are under threat by institutionalizing a system that is inclusive. More indigenous youths need to be allowed to participate in the preservation of their language and culture.

4.1 Recent Tasks in preservation

One of the first and most important contributions in codification of the Inpui language in Northeast India is the Bible Society of India's translation of the Bible into Inpui. And subsequently, the first Inpui dictionary was compiled by Inpui Students in New Delhi. These two accomplishments still remain like disparate monuments. Although in the recent, there had been an effort to publish a yearly magazine organized by the Inpui Students Union Delhi, its reach missed out a larger chunk of the population. The process of preservation needed institutionalization and incorporation of the indigenous language into the daily practices of the people- in their formal or informal, public and private spaces of interaction and communication.

5. Language Context and Role of Technology

Words and Language are context specific. A language loses its prominence when it loses its social and cultural context or is subdued by another new cultural and social change. The non-western ways of communication also face struggle due to the competition it faces from the ubiquity of modern technological social spaces and internet dominated by English, and uprooting of its language from its contextual specificity. But according to Mark Waschauer, internet is neither neutral nor deterministic, but rather a site of social struggle (Waschauer 1998). Yet definitely, the stakes are against the indigenous language especially in a post-colonial context like the Northeast India. But technology and internet can be shaped and made to be more accessible and friendlier to indigenous language through the bilingual mode of communication. The Hawaiian Leoki that provides a complex platform or services for indigenous speakers to engage electronically in their native language is a front runner in this revitalization endeavor (Warschauer et al., 1997) Among the indigenous tribe of Inpui or the larger Nagas, bilingual use in education is quite dismal. And even in spaces where formal education cannot reach, the internet space with its ubiquitous presence is an immensely important space for promotion of indigenous language.

6. Bilingual mode in Development

The importance of institutionalizing the bilingual use of communication is also deeply important in development. A lot of the government schemes are introduced in the dominant language of Hindi which is unintelligible to the indigenous communities. Hence the scheme remains secretive and inaccessible, as it is incomprehensible to the public. The bilingual nomenclature for these schemes can create ownership among the indigenous people and press the officials towards accountability towards the schemes. Thus, development is closely inter-woven with the use of the indigenous language.

[2] Brenzinger, et al. 2003

7. Conclusion

And thus, the bilingual opportunities for the indigenous people must be set in motion. The concerted efforts of the indigenous people with the experts, technocrats, policy makers and international organizations are of dire importance for the betterment of the indigenous peoples and the preservation of diversity in the world.

Bibliographical References

Bielenberg, Brian (1999): Indigenous Language Codification: Cultural Effects. In Jon Reyhner et al., editors "Revitalizing Indigenous Languages, pages 103-112, Flagstaff, Arizona, Northern Arizona University

Brenzinger, Matthias, Dwyer , Arienne M., Graaf, Tjeerd de, Grinevald ,Colette, Krauss , Michael, Miyaoka, Osahito, Ostler, Nicholas, Sakiyama, Osamu, Villalón, María E., Yamamoto, Akira Y. (2003): Language Vitality and Endangerment, Paris, UNESCO

Hornberger, Nancy H. (2012) Language Shift and Language Revitalization. Oxford Handbook Online. DOI: 10.1093/oxfordhb/9780195384253.013.0028

Khumba, K. Alung (2012) : The Kabui (Inpui) Naga : An Identity Crisis, Haochong, Manipur. Retrieved from https://drive.google.com/file/d/0B4uZqVViN7FOTNmNzY1ODctZmJiMC00ZWY0LWEzNzItZmMzYTYzMmIzYjM4/view

Mccarty, Teresa L. (2003): Revitalising Indigenous Languages in Homogenising Times. Comparative Education Volume 39 No. 2 2003, pp. 147-163

McLaughlin, Daniel (1995): Enabling bilingual program development in American Indian schools. The Bilingual Research Journal, Winter 1995, Vol. 19, No. 1, pp. 169-178

McLaughlin, D. (1992). When literacy empowers: Navajo language in print. Albuquerque, NM: University of New Mexico

Nancy H. Hornberger (2006) Voice and Biliteracy in Indigenous Language Revitalization: Contentious Educational Practices in Quechua, Guarani, and Māori Contexts, Journal of Language, Identity, and Education, 5:4, 277-292 DOI: 10.1207/s15327701jlie0504_2

Padel, Felix, (1996): The sacrifice of Human Being, Delhi, Oxford University Press.

Street, B. (1984). Literacy in theory and practice. Cambridge, UK: Cambridge University.

Warschauer, M. (1998): Technology and indigenous language revitalization: Analyzing the experience of

Hawai'i. Canadian Modern Language Review, 55(1), pp 140-161.

Warschauer, M., Donaghy, K., & Kuamoÿo, H. (1997). Leoki: A powerful voice of Hawaiian language revitalization. Computer Assisted Language Learning, 10(4), 349–361. doi:10.1080/0958822970100405

Acknowledgements

I would like to thank Charu Bikash – the AIPP Youth Capacity Building Coordinator and Susan Vize the Regional Advisor for Social and Human Science Asia Pacific Region from UNESCO Bangkok for nominating me to the LT4ALL Conference. I also like to convey my heartful thanks to David Young the Youth Programme Officer UNESCO Bangkok for the tedious feedback over this article. My gratitude extends to Sachoiba Inka- the editor of Kalakhwaanbang, the Inpui yearly magazine for his kind gesture and help offered to me. And I also convey my gratitude to NPMHR who had worked together, discuss and organize our issues of the indigenous people. And finally, I would like to pay my sincere gratitude towards the Organisers of LT4ALL for their tireless efforts in facilitating the conference and arranging the financial grant to make my participation at the conference possible. And I owe my love and gratitude to all the people who cared, loved and molded me and given me a sense of responsibility.

Proceedings, pages 193–197
Paris, UNESCO Headquarters, 5-6 December, 2019.

Language Shift, Language Technology, and Language Revitalization: Challenges and Possibilities for St. Lawrence Island Yupik

Lane Schwartz
University of Illinois at Urbana-Champaign
Department of Linguistics
lanes@illinois.edu

Abstract

St. Lawrence Island Yupik is a polysynthetic language indigenous to St. Lawrence Island, Alaska, and the Chukotka Peninsula of Russia. While the vast majority of St. Lawrence Islanders over the age of 40 are fluent L1 Yupik speakers, rapid language shift is underway among younger generations; language shift in Chukotka is even further advanced. This work presents a holistic proposal for language revitalization that takes into account numerous serious challenges, including the remote location of St. Lawrence Island and Chukotka, the high turnover rate among local teachers, socioeconomic challenges, and the lack of existing language learning materials.

Keywords: St. Lawrence Island Yupik, language technology, language revitalization

Piyuwhaaq

Sivuqam akuzipiga akuzitngi qerngughquteghllagluteng ayuqut, Sivuqametutlu pamanillu quteghllagmillu. Akuzipikayuget kiyang 40 year-eneng nuyekliigut, taawangiinaq sukallunteng allanun ulunun lliighaqut nuteghatlu taghnughhaatlu; wataqaaghaq pamani quteghllagmi. Una qepghaqaghqaq aaptaquq ulum uutghutelleghqaaneng piyaqnaghngaan uglalghii ilalluku uyavantulanga Sivuqaaamllu Quteghllagemllu, apeghtughistetlu mulungigatulangitnengllu, kiyaghtaallghemllu allangughnenganengllu, enkaam apeghtuusipagitellghanengllu.

1. The Inuit-Yupik language family

The Inuit-Yupik-Unganam Tunuu languages constitute the northernmost language family in the Western Hemisphere. Unganam Tunuu is indigenous to far southwest Alaska and the Aleutian Islands; the remaining languages constitute the Inuit-Yupik language family. The eastern branch of the Inuit-Yupik family, the Inuit languages, represent a dialect continuum indigenous to the Arctic and near-Arctic coast of North America, encompassing Greenland and the north coast of Canada and Alaska. The western branch of the family, the Yupik languages, are indigenous to western Alaska and the Bering Strait region, including St. Lawrence Island, Alaska and the Chukotka Peninsula of Russia. The Inuit-Yupik-Unganam Tunuu language family is notable as the only language family in the world indigenous to both North America and Asia (see Figure 1). This work will focus on St. Lawrence Island Yupik, a variety indigenous to St. Lawrence Island, Alaska and parts of the Bering Sea coast of Chukotka, Russia. St. Lawrence Island Yupik is the only language within the Inuit-Yupik family spoken natively on both continents.

1.1. Terminology

The Inuit-Yupik language family has historically been called the Eskimo language family. Within Alaska, the term Eskimo has been used to encompass the Inuit peoples of northern Alaska and the Yupik peoples of western and southwestern Alaska. Outside of Alaska, especially in Canada, the term Eskimo is now considered derogatory. In any case, the term is an exonym, and the Inuit Circumpolar Council has requested that its use be discontinued (ICCR, 2010). However, the use of the term Inuit to refer to Yupik peoples and languages obscures the fact that there is a historical and linguistic distinction between Inuit and Yupik. We therefore forgo the further use of the term Eskimo and instead use the term Inuit-Yupik to refer to the language family that encompasses the Inuit and Yupik languages, and the term Inuit-Yupik-Unganam (Fortescue et al., 2010) to refer to the broader family that also encompasses Unganam Tunuu (exonym: Aleut).[1]

The term Central Siberian Yupik was proposed in the 20th century to refer to the language spoken on St. Lawrence Island and across the Bering Strait in villages including Ungaziq, Chukotka (Russian name: Chaplino). The modifier term "Central" was chosen based on the idea that this language was centrally located amongst the three Yupik languages spoken in Chukotka both geographically and in terms of language relatedness (Michael Krauss, P.C. 2017). Subsequent research has made a strong case that Sirenik (one of the three) should be considered to form its own branch of the Inuit-Yupik family, rather than a branch of Yupik (see Figure 1). The modifier term "Siberian" was chosen to contrast with the Alaskan varieties of Yupik, with the term Siberia historically used in Alaska to refer to Chukotka. However, within Russia the term Siberia generally refers to a much broader region that mostly includes territory far west of Chukotka. In the interest of accuracy and clarity, we therefore follow Schwartz et al. (2020) in arguing that the term Central Siberian Yupik should be replaced by the term St. Lawrence Island Yupik in English and the term Chaplinski Yupik in Russian (refer-

[1] If a more concise name for the Inuit-Yupik-Unganam language family is at some point needed within the academic literature, the term *Iyut* could perhaps be proposed to the broader Inuit, Yupik, and Unganam communities. The term is an acronym formed from the language names (Inuit-Yupik-Unganam Tunuu) and has a surface form ending in *-t* which is consistent with how plural nouns are inflected in many of the languages in this family.

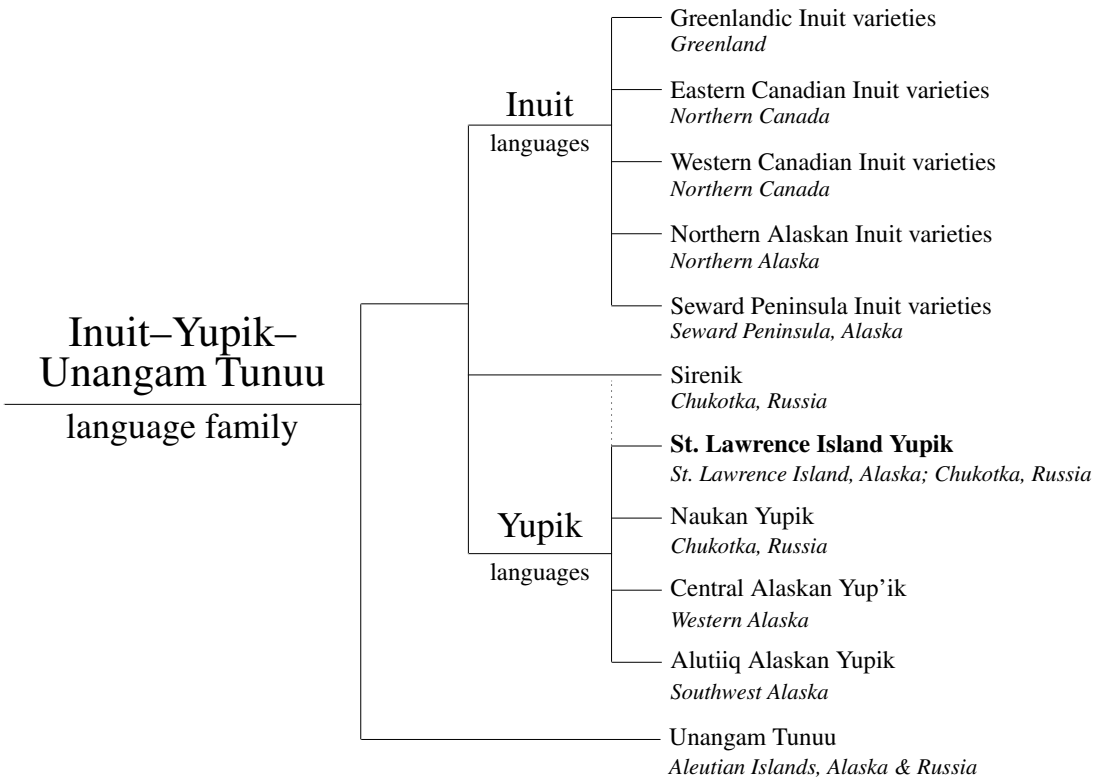


Figure 1: Inuit-Yupik-Unangam Tunuu language family (Fortescue et al., 2010; Krauss et al., 2011)

ring to Chaplino and New Chaplino, the respective Russian names of a prominent historical village and a contemporary Yupik village in Chukotka). The Yupik terms *Yupigestun* and *Akuzupik* are endonyms for this language, with specific terms such as *Sivuqaghhmistun* and *Ungazighhmistun* used to refer to the specific varieties spoken on St. Lawrence Island and Chukotka, respectively (Jacobson, 1990).

2. Status of St. Lawrence Island Yupik

St. Lawrence Island Yupik (ISO 639-3: *ess*) is a polysynthetic language indigenous to St. Lawrence Island, Alaska, and the Chukotka Peninsula of Russia. We use the term *Yupik* to refer to the language and to any individual Yupik person; we use the Yupik plural word *Yupiget* to refer to multiple Yupik persons. Schwartz et al. (2020) estimate 800–900 fully fluent L1 Yupik speakers out of an ethnic population of approximately 2400–2500 Yupiget. The majority of fluent Yupik speakers live in the villages of Gambell and Savoonga on St. Lawrence Island or in Chukotka, with smaller numbers residing in larger settlements, primarily Nome and Anchorage. While the vast majority of St. Lawrence Island Yupiget born in or prior to 1980 (individuals who as of late 2019 were 40 years old or older) are fluent L1 Yupik speakers (Krauss, 1980), rapid language shift is underway among younger generations; language shift among Yupiget in Chukotka is even further advanced (Morgounova, 2007). During linguistic fieldwork conducted over the period of 2016–2019, we observed very little Yupik use among Yupiget born prior to 1980, and widespread use of English among this age group.

2.1. Yupik Language in Education

Yupik language pedagogical materials were developed in the Soviet Union in the early 20th century (Krupnik and Chlenov, 2013) and in Alaska in the late 20th century (Koonooka, 2005). These materials were used in Russia into perhaps the 1950s (Krupnik and Chlenov, 2013) and on St. Lawrence Island until the early 2000s (Koonooka, 2005). The schools on St. Lawrence Island today have some basic Yupik language instruction, but very little of the pedagogical materials developed in the 1980s and 1990s are in use. The existing Yupik language books are archived at the Alaska Native Language Archive at the University of Alaska Fairbanks and the Materials Development Center in the Gambell School; these books are not easily accessible by members of the St. Lawrence Island community.

2.2. Challenges

The Yupik communities on St. Lawrence Island (and in Chukotka) are geographically very isolated. Like many indigenous communities, these communities have high rates of poverty, and many individuals within these communities struggle with various associated socioeconomic challenges. Within the school system, there is a chronic struggle to recruit and retain highly qualified teachers willing to serve these isolated communities. The certified teachers recruited to the schools are not Yupik speakers. A Yupik instructional curriculum was developed in the early 1990s, but it is in need of major pedagogical revision, and generally assumes a much higher level of Yupik proficiency among both instructors and students than is the case today. No pedagogi-

cal material designed to teach Yupik from scratch exists.

3. Yupik Language Technology

The polysynthetic nature of Yupik means that Yupik words are commonly composed of multiple morphemes. This means that the development of even simple language technologies such as spell-checkers and electronic dictionaries is substantially more complicated than for isolating or analytic languages. The first known publicly released Yupik language technology was released by Schwartz and Chen (2017) and included a simple web-based rule-based spell-checker that ensured that each word conformed to Yupik phonotactic and orthotactic requirements, as well as transliteration utilities allowing user-entered Yupik text to be transliterated into the Cyrillic orthography used in Chukotka, a Latin orthography used on St. Lawrence Island, a fully transparent alternative Latin orthography designed for pedagogical use, the Americanist phonetic notation used by several existing Yupik linguistic works, and the International Phonetic Alphabet notation. Chen and Schwartz (2018) implemented a prototype finite-state morphological analyzer for Yupik capable of decomposing a Yupik word into its constituent morphemes. Hunt et al. (2019) implemented a prototype web-based electronic dictionary for Yupik that incorporated a Javascript version of the Chen and Schwartz (2018) finite-state morphological analyzer. Schwartz et al. (2019a) explored the viability of using recurrent neural network methods to learn a more generalized morphological analyzer from the finite-state analyzer of Chen and Schwartz (2018). During the summer of 2019, a six-week summer research workshop (Schwartz et al., 2019c) explored this issue in substantially more depth, including the development of a language model prototype for use in text completion on mobile devices.

4. Goals and Proposals for Language Yupik Language Education and Revitalization

Over the course of our linguistic fieldwork conducted in Gambell between 2016 and 2019, a consistent theme voiced by many members of the Yupik community has been a desire for substantially strengthened Yupik instruction in the schools, ideally in the form of a Yupik language immersion program. This desire was communicated directly by the tribal council of the Native Village of Gambell, the elected representatives of the Yupik people in Gambell, during multiple meetings with the author of this work over the course of three years, as well as in conversations between this author and numerous members of the community. In addition to a strong Yupik-language educational program in the St. Lawrence Island schools, an additional goal stated by some members of the community is the establishment of self-study materials for use by younger Yupiget who are partially fluent, passively fluent, or not fluent in Yupik. Such materials would be especially helpful for younger adults of child-rearing age who are not themselves fluent Yupik speakers, but who have a desire to speak and transmit Yupik to their own children.

Current efforts in support of Yupik language education and revitalization efforts include the recent creation of a community-organized language committee on St. Lawrence Island (the Kaalguq Committee), language technology development led by the author of this work (see §3.), Yupik language documentation efforts led by Dr. Sylvia Schreiner of George Mason University and the author of this work, and ongoing efforts (begun in 2016) to digitize all printed Yupik-language materials into accessible PDF and plain-text formats.

Continued leadership from members of the St. Lawrence Island Yupik community is critical for the development and success of formal language revitalization efforts, as is consultation with stakeholders in the St. Lawrence Island community, including tribal and local governments, the Bering Strait School District, and the Alaska Native Language Center. The proposals in the following subsections (§4.1.–§4.3.) represent one possible forward path toward the long-term goal of a robust Yupik-language immersion program in the St. Lawrence Island schools.

4.1. Mobile-friendly digital Yupik library

During the 1970s through the early 1990s, a substantial number of Yupik-language materials were developed for educational use by the Nome Schools, the Bureau of Indian Affairs, the Alaska Native Language Center, and the Bering Strait School District. Since 2016, over ninety such books have been identified and scanned. This includes pre-primers for use in pre-school and early elementary settings, a set of three mid-elementary readers, and a three-volume set of stories by Yupik elders appropriate for use by advanced Yupik-language students.

Current efforts are underway to digitize all of these books into plain-text formats, and to obtain high-quality audio recordings of as many as possible. Once in plain-text format, each book (with associated audio if available) can be converted into an interactive e-book in e-pub format (Schwartz et al., 2019b). These e-books could then be gathered into a mobile-friendly collection, which community members could download for offline access when mobile internet services are active. Mobile phones are in widespread use on St. Lawrence Island, although mobile data access is relatively unreliable. By prioritizing books with the lowest reading level, this collection of audio e-books could be utilized by instructors in the St. Lawrence Island schools, as well as by adult language learners and their children at home.

4.2. Development of pedagogical materials

There are currently no Yupik self-study materials, no materials designed to teach Yupik from scratch in an educational setting, and no Yupik-language pedagogical materials for teaching other subjects (such as mathematics) using Yupik as the language of instruction. These facts represent major challenges to the eventual establishment of a Yupik language immersion school.

A medium-term goal, then, is the development of materials for use in learning Yupik. Ideally, such materials should be designed to be dual-use whenever possible, in order to support use in the St. Lawrence Island schools as well as use by adults in a self-study scenario. Materials should be designed keeping in mind that Yupik instruction in the

schools could be led by Yupik-dominant or English dominant teaching aides who may lack formal training, or potentially by (English-speaking) certified teachers from outside St. Lawrence Island who have formal training but not in language instruction.

Language technology could potentially be used to assist in the development of such pedagogical materials. Machine translation and translation memory technologies trained for translation from English (or from a different Inuit-Yupik language) into Yupik could potentially be used in a computer-aided translation scenario in which existing subject-matter textbooks are translated into Yupik. Yupik-language spell-checking and grammar modelling technology could also play a role.

4.3. Immersion programs

The academic literature on language revitalization (Hinton and Hale, 2001; Hinton, 2013) regarding best practices includes three related techniques that have been shown to be successful in creating new speakers of endangered languages: language nests (Hinton, 2013), language immersion programs (Hale, 2001), and master-apprentice programs (Hinton, 2001). In language nests, fluent adult speakers work with very young children in child-care or preschool-like settings. In school immersion programs, the immersion language is spoken by fluent instructors as the language of instruction in schools. In master-apprentice programs, an adult language learner or learners interact with a (usually elder) fluent speaker of the language in an immersion setting over an intense 2-3 year period, typically for 20–40 hours per week.

The establishment of a Yupik language immersion program in the St. Lawrence Island schools is a stated goal of many in the St. Lawrence Island community. Achieving this long-term goal will require coordination and buy-in from the Bering Strait School District, in addition to the development of Yupik-language subject-matter textbooks. An equally important requirement is the availability of instructors who can teach using Yupik as the language of instruction.

While it is theoretically possible that a sufficiently well-motivated student could learn Yupik through self-study (assuming that such self-study materials are first developed), the most promising path to the development of Yupik-fluent instructors is the establishment of a master-apprentice program. Full-time master-apprentice programs in which the master and the apprentices are paid a living wage for their time have been shown to be effective in developing new language speakers in a relatively short period of time (2–3 years). Community-led efforts to seek grant funding could provide one possible source for the establishment of such a master-apprentice program. Innovative partnerships between the Bering Strait School District, the St. Lawrence Island tribal governments, and the Alaska Native Language Center at the University of Alaska could also be explored, in order to create a novel program which could provide Yupik proficiency through a master-apprentice program while simultaneously providing students with some form of (possibly distance-based) teacher training.

5. Acknowledgements

I want to offer my immense thanks to the people of St. Lawrence Island for graciously sharing their language with me. The St. Lawrence Island Yupik language is a critical part of the cultural heritage of the St. Lawrence Island Yupik people. It has been and continues to be an immense honor to be welcomed into the St. Lawrence Island community, and to be trusted to engage in this work. I want to offer my great thanks to Petuwaq and all of the Yupik speakers who have worked with my colleagues and I since 2016. I want to offer my deep gratitude to the people of St. Lawrence Island, who first welcomed my family in 1982, and have generously continued to do so in the decades since. I want to thank the board members and staff of the Native Village of Gambell, the City of Gambell, and Sivuqaq, Inc, as well as the faculty and staff of Gambell Schools and the staff of Gambell Lodge.

I want to acknowledge the National Science Foundation and its Documenting Endangered Languages program, which has generously supported this research under NSF Award 1761680.

Finally, I want to offer my profound thanks to Willem de Reuse, Michael Krauss, and especially to Steve Jacobson for their decades-long work documenting the Yupik language. None of my work would have been possible without the outstanding grammar of Yupik that Dr. Jacobson developed over the course of many years working with the St. Lawrence Island community, as well as the incredible Yupik dictionary that he edited.

Igamsiqanaghhalek!

6. Bibliographical References

Chen, E. and Schwartz, L. (2018). A morphological analyzer for St. Lawrence Island / Central Siberian Yupik. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC'18)*, Miyazaki, Japan, May.

Fortescue, M., Jacobson, S., and Kaplan, L. (2010). *Comparative Eskimo Dictionary with Aleut Cognates*. Alaska Native Language Center, Fairbanks, Alaska, 2nd edition.

Hale, K. (2001). Linguistic aspects of language teaching and learning in immersion contexts. In Leanne Hinton et al., editors, *The Green Book of Language Revitalization in Practice*, page 227–235. Brill.

Leanne Hinton et al., editors. (2001). *The Green Book of Language Revitalization in Practice*. Brill.

Hinton, L. (2001). The Master-Apprentice language learning program. In Leanne Hinton et al., editors, *The Green Book of Language Revitalization in Practice*, page 217–226. Brill.

Leanne Hinton, editor. (2013). *Bringing our Languages Home: Language Revitalization for Families*. Heyday, Berkeley, California.

Hunt, B., Chen, E., Schreiner, S. L., and Schwartz, L. (2019). Community lexical access for an endangered polysynthetic language: An electronic dictionary for St. Lawrence Island Yupik. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*,

pages 122–126, Minneapolis, Minnesota, June. Association for Computational Linguistics.

ICCR. (2010). Inuit circumpolar council resolution 2010–01 on the use of the term inuit in scientific and other circles.

Jacobson, S. A. (1990). *A Practical Grammar of the St. Lawrence Island / Siberian Yupik Eskimo Language, preliminary edition.* Alaska Native Language Center, Fairbanks, Alaska.

Koonooka, (Petuwaq), C. (2005). Yupik language instruction in Gambell (St. Lawrence Island, Alaska). *Études/Inuit/Studies*, 29(1/2):251–266.

Krauss, M., Holton, G., Kerr, J., and West, C. T. (2011). Indigenous peoples and languages of Alaska. ANLC Identifier G961K2010.

Krauss, M. (1980). Alaska Native languages: Past, present and future. *ANLC Research Papers*, 4.

Krupnik, I. and Chlenov, M. (2013). *Yupik Transitions — Change and Survival at Bering Strait, 1900-1960*. University of Alaska Press, Fairbanks, Alaska.

Morgounova, D. (2007). Language, identities and ideologies of the past and present Chukotka. *Études/Inuit/Studies*, 31(1-2):183–200.

Schwartz, L. and Chen, E. (2017). Liinnaqumalghiit: A web-based tool for addressing orthographic transparency in St. Lawrence Island/Central Siberian Yupik. *Language Documentation and Conservation*, 11:275–288, September.

Schwartz, L., Chen, E., Hunt, B., and Schreiner, S. L. (2019a). Bootstrapping a neural morphological analyzer for St. Lawrence Island Yupik from a finite-state transducer. In *Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 87–96, Honolulu, February. Association for Computational Linguistics.

Schwartz, L., Schreiner, S. L., Zukerman, P., Soldati, G. M., Chen, E., and Hunt, B. (2019b). Initiating a tool-building infrastructure for the use of the St. Lawrence Island Yupik language community. International Year of Indigenous Languages 2019: Perspectives Conference, October.

Schwartz, L., Tyers, F., Park, H., Steimel, K., Strunk, L., Haley, C., Zhang, K., Kirov, C., Knowles, R., Levin, L., Littell, P., Lo, J., Prud'hommeaux, E., and Jimmerson, R. (2019c). 2019 JSALT Workshop on Neural Polysynthetic Language Modelling, June-August.

Schwartz, L., Schreiner, S., and Chen, E. (2020). Community-focused language documentation in support of language education and revitalization for St. Lawrence Island Yupik. *Études/Inuit/Studies*, In press.

Schwartz, L. (2018). NNA: Collaborative research: Integrating language documentation and computational tools for Yupik, an Alaska Native language. U.S. National Science Foundation Award 1761680.

Proceedings, pages 198–201
Paris, UNESCO Headquarters, 5-6 December, 2019.

Copyright in the context of tooling up Corsican and other less-resourced languages

Laurent Kevers, Stella Retali-Medori
UMR CNRS 6240 LISA, Università di Corsica - Pasquale Paoli
Avenue Jean Nicoli, 20250 Corte, France
{kevers_l, medori_e}@univ-corse.fr

Abstract

Anyone trying to gather linguistic resources for Natural Language Processing (NLP) will sooner or later be facing the legal aspects, mainly related to copyright, that arise from this activity. These difficulties often occur when collecting corpora, which is generally among the top priorities for processing less-resourced languages. While the current legislative framework is not adequate, it seems that positive developments are emerging. Various actions can also be considered to support this evolution.

Keywords: less-resourced languages, corpora, linguistic resources, copyright, Corsican language

Résumé

Toute personne qui essaye de rassembler des ressources linguistiques pour le Traitement Automatique d'une Langue (TAL) sera tôt ou tard confrontée aux aspects légaux, principalement liés au droit d'auteur, que soulève cette activité. Ces difficultés se matérialisent souvent lors de la collecte de corpus, qui se situe généralement parmi les premières priorités pour le traitement des langues peu dotées. Si le cadre législatif actuel n'est effectivement pas adapté, il semble que des évolutions positives se profilent. Différentes actions peuvent aussi être envisagées pour accompagner ce changement.

1. The Corsican language, a less-resourced language

Corsican is a Latin language and is part of the Italo-Romance domain. It has known various contacts and linguistic influences. From a dialectal point of view, four or even five areas are identifiable (Dalbera-Stefanaggi, 2002; Dalbera-Stefanaggi, 2007), but they constitute a *continuum* and do not prevent interunderstanding between speakers. The spelling of Corsican is, with some adaptations, based on the Italian graphic system[1]. However, despite the implementation of a polynomic approach (Marcellesi, 1984) that encompasses all dialectal variants, the writing of the language is not standardized.

Nowadays, Corsican is, with French, part of a diglossic language environment, and its use is declining. The development of tools is necessary for its preservation, enhancement, transmission and promotion[2]. A policy in the service of the Corsican language is active on the island territory, in particular for its development through new technologies. However, if several tools and works exist for the learning and linguistic description of Corsican dialects, their inclusion in the digital humanities domain remains insufficient.

In particular, sites and applications dedicated to translation, lexicon and syntax contain little data in comparison with the richness and complexity of the language. On the other hand, this wealth is found on databases such as the *Banque de Données Langue Corse* (BDLC) and *Infcor*[3] (*Banca di dati di a lingua corsa*). To our knowledge, there are very few resources and tools designed for Natural Language Processing (NLP) in Corsican. The ELDA 2014 report on linguistic resources dedicated to the languages of France (Leixa et al., 2014) lists 93 resources for Corsican. More than a third of these are recordings and transcriptions from the BDLC project. Corsican therefore falls into the category of less-resourced languages.

2. Development of NLP resources and tools for Corsican

Given this observation, we have decided to work to improve the situation of the Corsican language with regard to its place in the digital world, and more particularly in the field of Natural Language Processing. To achieve this objective and start tooling up the Corsican language, we rely on the BDLC project. This project[4] is designed in a scientific context and hosts linguistic data related to Corsican know-how and cultural traditions throughout the island territory. It is naturally linked to the *Nouvel Atlas Linguistique et ethnographique de la Corse* (NALC).

We have defined a roadmap outlining the actions to be undertaken (Kevers et al., 2019). These are generally in line with those put forward by Ceberio Berger et al. (2018). We started by collecting corpora and setting up an online consultation interface in the form of a concordancer, implementing a language detection tool and building an electronic dictionary. In the long term, we plan to work on a part-of-speech tagger.

The question we want to highlight in this article concerns the legal aspects, mainly related to copyright, that any person trying to gather linguistic resources for a language inevitably encounters. These difficulties often materialize

[1] See Retali-Medori (2015)

[2] According to the recommendations of UNESCO Ad Hoc Expert Group on Endangered Languages (2003)

[3] `http://infcor.adecec.net`

[4] See `http://bdlc.univ-corse.fr`. A synthesis of the project history is presented by (Dalbera-Stefanaggi and Retali-Medori, 2015).

when collecting corpora, which is generally among the top priorities for processing less-resourced languages.

3. Legal aspects : corpora and copyright

3.1. Introduction

In addition to documenting the language, there are many uses for corpora, starting with comparing the intuition and linguistic knowledge of language specialists with large "real" datasets. Corpora can also be useful for building lexical resources, for creating automatic processing tools, especially through machine learning, or even in the educational field.

This task faces two main obstacles: the availability of documents, preferably in a digital form, and their legal terms of use.

Apart from the question of the existence of the documents, the first difficulty is essentially technical. The first step is to identify existing resources and process them according to their nature. In the case of printed documents, it will be required to digitize them. If they are already in a digital format, conversion operations[5] or even "harvesting"[6] may be necessary.

The second difficulty lies in respecting the rights that apply to this content. Indeed, the copyright laws do not generally allow their free and complete use, even for research purposes. This obstacle constitutes a real limitation for research in general, and for the digital development of less-resourced languages in particular, and has therefore been highlighted on many occasions, including by Zayed et al. (2016) : *One of the big obstacles for the current research is the lack of large-scale freely-licensed heterogeneous corpora in multiple languages, which can be redistributed in the form of entire documents. [...] due to the restrictive license of the content, many corpora cannot be re-distributed because of the risk of copyright infringement*.

The task of automatic corpora building from the web[7] is particularly affected by this problem. Tools proposed for this purpose, such as BootCaT[8] (Baroni and Bernardini, 2004) or Sketch Engine[9] (Kilgarriff et al., 2014), will be difficult to use if it is planned to redistribute the resources and tools created from these corpora.

3.2. Current situation

The legal analyses that are reproduced later in this article come mainly from Geiger et al. (2019). Our objective here is to summarize them by highlighting the main points, so that actors in the world of Text and Data Mining (TDM) who are not aware of the issues, can take note of it.

TDM research is faced with a legal situation that does not allow it to proceed smoothly, given its potential involvement in copyright issues : *[...] during the chain of activities enabling TDM research, technically some IPR relevant actions are necessary so that in the absence of a specific permission within the legal framework, TDM can lead to an infringement*, (Geiger et al., 2019, p.7). In particular, copying and modification of copyrighted works may be problematic : *TDM usually involves some copying, which even in case of limited excerpt might infringe the right of reproduction. [...] any reproductions resulting in the creation of a copy of a protected work along the chain of TDM activities might trigger copyright infringement. In this respect, preprocessing to standardize materials into machine-readable formats might trigger infringement of the right of reproduction*, (Geiger et al., 2019, p.7-8). In addition to textual documents, these limitations also apply to the database protected by the *sui generis* right.

This situation also poses a problem in terms of scientific approach : *[...] contemporary research practices, striving for verifiability of TDM research results, require the ability of researchers to store source materials and to communicate them at least to their peers. From a legal perspective, this conduct could most likely trigger the infringement of the right of communication to the public*, (Geiger et al., 2019, p.9). Similarly, the diffusion of models learned and derived from non-free sources, which constitute a transformed state of the original work, places researchers in a legitimate position of uncertainty about the legal implications of their work. The only elements that would be risk-free to communicate would be the final results produced by the TDM procedure : *it is to be noted that the TDM output should not infringe any exclusive rights as it merely reports on the results of the TDM quantitative analysis, typically not including parts or extracts of the mined materials*, (Geiger et al., 2019, p.9).

Even if some exceptions exist and can be used, the current European legal framework does not allow the development of TDM projects in a serene manner : *All in all, the possibility of relying on existing provisions — including temporary acts of reproduction, scientific research, private use, normal use of a database, and extraction of "insubstantial parts" from a database protected by the sui generis right — without adoption of additional interpretative norms or judgements of high instances was doubtful*, (Geiger et al., 2019, p.17).

Finally, it should be noted that various European countries have taken initiatives to develop copyright exceptions for TDM. For example, France allows the : *reproduction from "lawful sources" (materials lawfully made available with the consent of the rightholders) for TDM as well as storage and communication of files created in the course of TDM research activities*, (Geiger et al., 2019, p.25). However, this exception to copyright is only granted if it occurs as a part of a scientific writing. In general, without going into the details about the exceptions provided by the various countries, these legal developments ultimately retain a certain degree of uncertainty as to their real ability to meet the legal needs faced by the TDM community. Moreover, the lack of a uniform approach should be underlined.

[5] Such as switching from PDF to text format.

[6] For content published in the form of websites.

[7] An ACL *Special Interest Group* (SIG) is dedicated to this domain under the name of *Web AS Corpus* (SIGWAC - https://www.sigwac.org.uk/).

[8] https://bootcat.dipintra.it/

[9] https://www.sketchengine.eu/

3.3. Future developments

However, Directive 2019/790/EU of the European Parliament and of the Council on copyright and related rights in the Digital Single Market[10], adopted on 17 April 2019, should improve the situation.

Indeed, this text introduces new exceptions to copyright, in particular *for reproductions and extractions made by research organisations and cultural heritage institutions in order to carry out, for the purposes of scientific research, text and data mining of works or other subject matter to which they have lawful access* (article 3, paragraph 1).

In addition to this exception specific to the field of scientific research, a more general exception is also provided (article 4). However, there is a restriction on the latter which limits its application to cases where the works concerned *has not been expressly reserved by their rightholders in an appropriate manner, such as machine-readable means in the case of content made publicly available online* (article 4, paragraph 3).

It should be noted that this directive must be transposed into the national laws of the Member States in order to be implemented, which should be the case by 2021.

4. Discussion

In light of the above, we wanted to identify some points for which actions can be undertaken.

- In the meantime, until a perfectly adapted legal framework is in place, we think it is necessary to initiate or continue work to provide adequate resources for TDM research. In this respect, the use of licenses that allow certain exceptions to copyright — for example the Creative Commons[11] family — should be encouraged and intensified. Obviously, the dialogue with the rights holders is a complex, sometimes time-consuming task that goes beyond the strict framework of research activities, but worth leading.
- It could be helpful to take initiatives to make legal information on copyright issues more visible and more easily accessible to a research audience which, while generally of good will, does not necessarily give all the necessary attention to these questions. In a sense, this article is a modest contribution to this goal. Our wish would be that this question could be taken into account in a more in-depth manner, by teams of both lawyers and researchers, and that it gives rise to a wider dissemination.
- Finally, it also seems useful to raise awareness among the legislative bodies in order to change the legal framework as quickly as possible, in particular with regard to the transposition of the Directive.

5. Bibliographical References

Baroni, M. and Bernardini, S. (2004). BootCaT: Bootstrapping Corpora and Terms from the Web. In *Proceedings of the fourth international conference on Language Resources and Evaluation (LREC 2004)*.

Ceberio Berger, K., Gurrutxaga Hernaiz, A., Baroni, P., Hicks, D., Kruse, E., Quochi, V., Russo, I., Salonen, T., Sarhimaa, A., and Soria, C. (2018). *Digital Language Survival Kit. The DLDP Recommendations to Improve Digital Vitality*. The Digital Language Diversity Project. Available at `http://www.dldp.eu/sites/default/files/documents/DLDP_Digital-Language-Survival-Kit.pdf`.

Dalbera-Stefanaggi, M.-J. and Retali-Medori, S. (2015). Trente ans de dialectologie corse : autour du programme Nouvel Atlas Linguistique et Ethnographique de la Corse et Banque de Données Langue Corse. In Stella Retali-Medori, editor, *Actes du colloque Tribune des chercheurs, études en linguistique*, volume 6 of *Corse d'hier et de demain - Nouvelle série*, pages 17–25, Bastia, France, June. Société des Sciences Historiques et Naturelles de la Corse.

Dalbera-Stefanaggi, M.-J. (2002). *La langue corse*. Number 3641 in Que sais-je? PUF, Paris, June.

Dalbera-Stefanaggi, M.-J. (2007). *Nouvel atlas linguistique et ethnographique de la Corse : Volume 1, Aréologie phonétique, édition revue et corrigée*. Comité des travaux historiques et scientifiques - CTHS, Ajaccio : Paris, Alain Piazzola edition, December.

Geiger, C., Frosio, G., and Bulayenko, O. (2019). Text and Data Mining: Articles 3 and 4 of the Directive 2019/790/EU. Research Paper No. 2019-08, Centre for International Intellectual Property Studies (CEIPI).

Kevers, L., Guéniot, F., Tognotti, A. G., and Retali-Medori, S. (2019). Outiller une langue peu dotée grâce au TALN : l'exemple du corse et de la BDLC. In *Actes de la 26e conférence sur le Traitement automatique des langues naturelles (TALN)*, Toulouse, France, July.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., and Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1(1):7–36, July.

Leixa, J., Mapelli, V., and Choukri, K. (2014). *Inventaire des ressources linguistiques des langues de France*. ELDA, September. Available at `http://www.elda.org/media/filer_public/2014/12/17/rapport_dglflf_05112014-1.pdf`.

Marcellesi, J.-B. (1984). La définition des langues en domaine roman : les enseignements à tirer de la situation corse. In *Actes du Congrès de Linguistique et de Philologie Romanes 5*, pages 307–314, Aix-en-Provence.

Retali-Medori, S. (2015). La documentation corse. In Eugeen Roegiest Maria Iliescu, editor, *Anthologies, textes, corpus et sources des langues romanes*, number 7 in Manuals of Romance Linguistics, pages 558–564. De Gruyter, Tübingen.

UNESCO Ad Hoc Expert Group on Endangered Languages. (2003). *Language Vitality and Endangerment. Document submitted to the International Expert Meeting on UNESCO Programme Safeguarding of Endangered Languages*. United Nations Educational, Scientific and Cultural Organization, Paris. Available at `https://ich.unesco.org/doc/src/00120-EN.pdf`.

Zayed, O., Habernal, I., and Gurevych, I. (2016).

[10] `http://data.europa.eu/eli/dir/2019/790/oj`

[11] `https://creativecommons.org/licenses/`

C4corpus: Multilingual Web-size Corpus with Free License. In Nicoletta Calzolari, et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, May.

Proceedings, pages 202–205
Paris, UNESCO Headquarters, 5-6 December, 2019.

Formal Models and Software Tools for the Computer Processing of the Tatar Language

Suleymanov D., Khusainov A., Gilmullin R.
Institute of Applied Semiotics of the Tatarstan Academy of Sciences,
Kazan Federal University
Kazan, Russia
{dvdt.slt, khusainov.aidar, rinatgilmullin}@gmail.com

Abstract

The paper describes models and software tools developed mainly as part of the State program for the preservation, study and development of the state languages of the Republic of Tatarstan and other languages in the Republic of Tatarstan. The content of the paper shows the current state of work on creating software tools and systems to support the Tatar language in computer technologies. The introduction provides a brief chronological summary of the results of research and development of the Institute of Applied Semiotics of the Tatarstan Academy of Sciences on the national computer systems localization and the use of the Tatar language in Information Technology, showing their compliance with modern trends in the development of natural language processing technology.

Keywords: Tatar language, language resources, NLP tools

Абстракт

Мәкаләдә Татарстан Республикасы дәүләт телләрен һәм Татарстан Республикасындагы башка телләрне саклау, өйрәнү һәм үстерү буенча Татарстан Республикасы дәүләт программасын гамәлгә ашыру кысаларында эшләнгән модельләр һәм программалар тасвирлана. Мәкаләнең эчтәлеге компьютер технологияләрен татар теле белән тәэмин итү өчен төзелгән программалар һәм системаларның бүгенге торышын чагылдыра. Керештә Татарстан Республикасы Фәннәр академиясе гамәли семиотика институтының компьютерларны татарчалаштыру һәм татар телен инфокоммуникация технологияләрендә куллану юнәлешендәге тикшеренүләре һәм эшләнмәләре турында аңлатма бирелә, аларның заманча технологияләр нигезендә башкарылыуы күрсәтелә.

1. Introduction

Natural Language Processing (NLP) is the general direction of artificial intelligence and mathematical linguistics. Currently, for the most spoken languages of the world, specialized software tools have been developed to support national languages in information technology.

On the other hand, there is a class of low-resourced languages that suffer from a lack of available language resources and software. Of particular interest to the Republic of Tatarstan are technologies that support the Turkic languages, including the Tatar language that is also khown as a low-resourced language.

The work in the field of Tatar NLP began in early 1990s in Tatarstan Academy of Sciences and Kazan State University. In this paper we present the main results obtained during this time, as well as basic information about the Tatar language and the short summary of the development history.

1.1 The Tatar Language

Tatar is the second spoken language in Russia. There are 4.2 million of speakers in Russia and near 5 million of speakers in the world (Eberhard et al., 2019). The Cyrillic alphabet (unified in 1939) consists of 39 characters. There are 12 vowel and 28 consonant sounds. Different dialects of Tatar can be identified: Western, Kazan (Middle) and Eastern. Based on the existing language classification (Berment, 2004; Krauwer, 2003), it was assigned to the under-resourced language class (Khusainov, 2014). However, recent results in machine translation, speech analysis and synthesis fields can change this situation.

1.2 History

Research and development in the field of computer linguistics for Tatar began in 1993 as part of the Joint Laboratory for Artificial Intelligence of the Academy of Sciences of the Tatarstan Republic and Kazan State University, which in 2009 was transformed into the Institute of Applied Semiotics of the Academy of Sciences of the Republic of Tatarstan.

All these years, one of the most important scientific and applied problems has been the development of software tools and linguistic resources for the widespread use of the Tatar language in information technology, including mobile devices and the Internet. The particular importance of the Tatar localization of computer technology is also determined by the need to ensure the parity functioning of the Tatar language along with the Russian language as the state language in the Republic of Tatarstan.

Research and development in computer technology for the Tatar language began in the end of the 1980s from the implementing of the first monitor and printer drivers, a text editor and the Tatar spellchecker, localization of computer publishing systems needed for the publication of Tatar books, newspapers and magazines.

It is interesting to look at the chronology of early research and development:

– 1992 year. Multimedia compact discs with the Tatar language teaching program “My First Tatar Dictionary”,

as well as the educational multimedia CD “Tatar Telle Zaman”, which includes, in addition to the training, testing, and dialogue blocks, also a block of linguistic games. All implementations provide training in trilingual (Tatar, Russian, English) mode. Moreover, the Tatar language operates in two modes: on the basis of the Cyrillic alphabet and on the basis of the Latin alphabet.
– 1994 year. The concept and architecture of the machine fund of the Tatar language were developed and the filling of the machine fund with electronic dictionaries and texts, with modules for processing Tatar texts, data processing systems in the Tatar language was started.
– 1995 year. A spellchecker for Tatar texts has been developed, which allows to find and correct spelling errors in Tatar texts.
– 1996 year. Joint with ABBYY company, a version of the FineReader OCR program was developed for Tatar.
– 1998 year. Joint with the Belkent University (Turkey), a Tatar two-level analyzer was developed that performs morphological analysis and generation of Tatar word forms.
– 2000 year. A diphone-based Tatar speech synthesizer was developed.
– 2003 year. A complete structural and functional model of the Tatar morphology was created.

2. Tatar Localization

National localization of computer systems is very important for the preservation and development of low-resource languages, as well as a necessary base for the development of language resources, natural language processors and programs for e-learning.
The most important and knowledge-intensive step in localization is the development and adoption of standards for knowledge representation and of a terminology for computer science and information technology.
The main criteria for Tatar localization are correctness, accuracy of the translation of the text, its semantic correspondence to the original text; brevity and clarity of texts (instructions, actions) on the menu buttons; clarity and compactness of texts in reference files. Typical derivational and syntactic models were developed for the different screen components, and the overall style of the interface was created on this basis.
In order to create and use linguistic resources for low-resource languages in all spheres of their manifestation (such as science, education, publishing, information processing etc.), it is important to integrate knowledge and skills of experts in related spheres: computer science, mathematics and linguistics.
Currently, the following important linguistic resources and software have been developed and are actively used while being under further development. Thanks to the cooperation of the Academy of Sciences of the Tatarstan Republic with Microsoft company, all versions of the OS Windows, beginning from the Windows NT, were localized.
The Tatar language became the second Turkic language after the Turkish language, localized by specialists of the republic itself, and not by the developers of the company. Microsoft Office applications were also localized, including the interface and help files, as well as the Tatar spellchecker.

Currently, the Tatar language is being actively introduced into mobile devices. Localized service applications are developed: keyboards, dictionaries, predictive typing systems, games, tutorials. In 2016, Tatar localization of the Russian mobile operating system Aurora OS began (jointly with the “Open Mobile Platform” company). Aurora OS has become the first mobile operating system that makes it possible to fully use the Tatar language along with Russian in mobile devices.
Another unique software system adapted for the Tatar language by our specialists together with «ABBYY LS» company is SmartCAT professional translation system (http://smartcat.ai/). The system is intended for widespread use as a tool for a professional translator with various useful functions (machine translation, electronic dictionaries, etc.). This platform for automation of translation, which optimizes the work process and comprehensively solves all translation tasks, allowing you to create projects, monitor the work of the translation team in real time, check translated segments, discuss details with the team directly in the system. The SmartCAT system is currently implemented in organizations and departments in all regions of the Republic of Tatarstan.

3. Software and Linguistic Resources for Tatar

The development of software tools, applications, and linguistic resources for the Tatar language ensures the use of computer systems and technologies for working with the Tatar language in all spheres and forms of its manifestation.

3.1 Tatar National Corpus “Tugan Tel”

The Tatar corpus "Tugan Tel" is a linguistic resource of the modern literary Tatar language. The corpus is addressed to a wide range of users: linguists, specialists in the field of Tatar, Turkic and general linguistics, typologists, teachers of the Tatar language, as well as everyone who studies and is interested in the Tatar language.
The volume of the corpus is over 200 million word forms and contains texts of various genres (fiction, media texts, texts of official documents, textbooks, scientific publications, etc.). Each document has a meta description (authors, their gender, creation dates, genres, parts, chapters, etc.). The texts included in the corpus are provided with morphological annotation (information about the part of speech and the grammatical characteristics of the word form). Morphological annotation of corpus texts is carried out automatically using the module of two-level morphological analysis of the Tatar language, implemented via PC-KIMMO software toolkit.

A search system has been developed for the corpus that allows searching for material by lexeme, word form, as well as by individual grammatical characteristics.
Corpus is available at https://tugantel.tatar.

3.2 Electronic Version of the Atlas of Tatar Dialects

In 2011-2012, an electronic version of the atlas of Tatar dialects was created. The Atlas includes all the main areas of Tatars settlement and reflects information on the phonetics, morphology, vocabulary and syntax of the Tatar language, collected in 28 regions of Russia.
The electronic Atlas database contains information on the distribution of the meanings of 215 linguistic phenomena across 1047 settlements. Maps display the features of Tatar dialects in the phonetics (68 cards), morphology (49 cards), vocabulary (93 cards) and syntax (5 cards) sections. The maps of the Atlas provide information on the distribution of dialects in the selected settlements.
The release of the electronic version of the atlas is a new stage in the presentation of dialectical knowledge of the Tatar language based on geographic information systems. An electronic Atlas is available at http://atlas.antat.ru/.

3.3 Morphological analysis and disambiguation

The Tatar language has a rich and regular, almost automatic, morphology (Suleymanov, 1998). The morphological model of the Tatar language is a basic component in almost all fully functional linguistic analyzers. Accordingly, the creation of a computer model of the Tatar morphology was one of the first and important tasks. Given the structural specificity of the Tatar language and based on applied problems, three different morphological models have been developed to date.
Generative morphology model based on affixing rules, although inferior to other models in speed, provides the completeness of the analysis of the word form, allowing you to fully take into account the agglutinative nature of the language, recognizing word forms of potentially unlimited length.
The paradigmatic model of Tatar morphology provides quick recognition of word forms and analysis of the correctness of Tatar word forms with an accuracy of 95% and is used in MS Windows and its Office applications. In addition, in a joint project with the Belkent University (Turkey), a two-level model of the Tatar language morphology was developed, implemented via the PC KIMMO software shell. A hybrid model of morphological analysis has also been created, using generative and paradigmatic approaches, which is part of the “Tatar morpheme” information system.

3.4 Tatar Speech Recognition and Synthesis

Systems of automatic recognition of continuous speech and its synthesis allow to carry out work on the implementation of human-machine speech interface.
A set of speech technologies is being developed at the Institute of Applied Semiotics, which includes the ability to identify the language of the speaker, automatic speech recognition and synthesis for Tatar. Databases of textual and speech information in the Tatar language are accumulated and analyzed, machine learning technologies are developed, the speech interface in the Tatar language is integrated into modern PCs and mobile devices.
Further development of speech technologies will open up prospects for sharing research results in the field of semantic analysis of text in the Tatar language, and will allow the creation of intelligent systems for helping the visually impaired, speech translators, intellectual assistants, etc.

3.5 Russian-Tatar Neural Machine Translation

Russian and Tatar languages are the official languages in the Republic of Tatarstan. This fact makes urgent the task of providing the population, state and other institutions with the possibility of automatic translation between these languages.
One of the key areas of activity of the Institute of Applied Semiotics of the Tatarstan Academy of Sciences is the creation of a machine translator in a Russian-Tatar language pair. An important component of the machine translator is linguistic support, which currently includes 1 million Russian-Tatar pairs. The created version of the Russian-Tatar translator (https://translate.tatar) is currently the best among the analogues in terms of translation quality.
The results of constructing a machine translator system for the Russian-Tatar language pair show that modern neural network algorithms and approaches are able to solve the translation problem at a fairly high level even for low-resourced language pairs.

4. Conclusion

The article describes the models and software tools developed for the Tatar language.
The description reflects the current state of work on the development of software tools and systems to support the Tatar language in computer technologies and includes a description of the morphological analyzer of the Tatar language, the system of machine translation for the Russian-Tatar language pair, the synthesizer and recognizer of Tatar speech, the electronic corpus of the Tatar language "Tugan tel”, as well as the electronic Atlas of Tatar dialects. We also give an overview of another important part of research and work on the localization of software products: operating systems (including mobile), mobile applications, websites.

5. Bibliographical References

Eberhard, David M., Gary F. Simons, and Charles D. Fennig (eds.). 2019. Ethnologue: Languages of the World. Twenty-second edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.

V. Berment, “Me´thodes pour informatiser des langues et des groups de langues peu dote´es”, Ph.D. Thesis, J. Fourier University, Grenoble I, 2004.

S. Krauwer, “The basic language resource kit (BLARK) as the first milestone for the language resources

roadmap", In Proc. of International Workshop Speech and Computer SPEECOM, Moscow, Russia, 2003, P. 8–15.
A. Khusainov, "Tekhnologiya avtomatizatsii sozdaniya I otsenki kachestva programmnikh sredstv analiza rechi c uchetom osobennostey maloresursnykh yazikov", Ph.D. Thesis, Kazan, 2014, 162 p.
D. Suleymanov, "Formalnaya elegantnost I estestvennaya slozhnost morfologii tatarskogo yazyka", In. Proc. of Information Technology in the Humanities, Kazan, Russia, 1998. URL: http://www.kcn.ru/_tat_ru/universitet/gum_konf/ot7.htm.

Proceedings, pages 206–209
Paris, UNESCO Headquarters, 5-6 December, 2019.

Indonesian Phoneme Set, Vocabulary, and Pronunciation for Automatic Speech Recognition and Speech Synthesizer

Dessi Puji Lestari[1,2], Roland Hartanto[1], Devin Hoesen[2], Guntario Sukma Cahyani[2], Sakriani Sakti[3§]

[1]Institut Teknologi Bandung, Jl. Ganesha 10 Bandung, Indonesia
[2]Prosa.ai, Jl. dr. Otten 10 Bandung, Indonesia
[3]Nara Institute of Science and Technology, 8916-5 Takayama, Nara 630-0192, Japan
dessipuji@stei.itb.ac.id, 13515107@std.stei.itb.ac.id, {devin.hoesen, guntario.cahyani}@prosa.ai, ssakti@is.naist.jp

Abstract

This paper describes the design of the Indonesian phoneme set, vocabulary and pronunciation applied for two main speech technology applications, automatic speech recognition and automatic speech synthesizer research for Bahasa Indonesia. There are 32 standard Indonesian phonemes, but they do not include variations in sound pronunciation or allophones caused by dialect and foreign language influences. Some studies of the Indonesian speech recognition systems even try to reduce the standard phoneme class, especially sounds that caused by the influence of foreign languages that are difficult to pronounce by Indonesians and diphthong sounds to improve the accuracy of speech recognition systems. On the other hand, from the results of the speech synthesizing system it was found that 32 standard Indonesian phonemes without incorporating allophones were insufficient to represent the pronunciation of Indonesian, thus the synthesized speech was not as natural as the original speaker.

Keywords: Bahasa Indonesia, phoneme, vocabulary, pronunciation, speech recognition, speech synthesizer

Abstrak

Pada makalah ini dipaparkan rancangan set bunyi atau fonem, kosakata, dan pelafalan kata yang digunakan pada penelitian-penelitian sistem pengenal ucapan dan sistem pensintesis ucapan untuk Bahasa Indonesia. Terdapat 32 bunyi baku bahasa Indonesia, namun bunyi-bunyi tersebut belum mencakup variasi-variasi pelafalan bunyi atau alofon yang disebabkan oleh pengaruh dialek dan pengaruh bahasa asing. Beberapa penelitian sistem pengenal ucapan bahasa Indonesia bahkan berusaha mengurangi kelas bunyi standard tersebut terutama bunyi-bunyi yang disebabkan oleh pengaruh bahasa asing yang sulit diucapkan oleh orang Indonesia dan bunyi diftong untuk meningkatkan akurasi sistem pengenal ucapan. Sebaliknya, dari hasil penelitian sistem penyintesis ucapan didapat bahwa 32 bunyi baku bahasa Indonesia tanpa memasukkan alofon dinilai masih kurang mewakili pelafalan bahasa Indonesia, sehingga suara hasil sintesis terdengar tidak sealami pembicara aslinya.

1. Introduction

Malay language is the root of Indonesian language called Bahasa Indonesia. For a long time, Malay language had indeed been used as an intermediary language or social language in Indonesia archipelago, Brunei, and Malaysia. This language spread very quickly in almost all regions in Indonesia since the 7th century and even formed a separate language variant that differed from its root, called "van Ophuijsen Malay" (Hasan, 1999). This language became widely known among indigenous people and was to be the national identity of Indonesia. It was then formalized as the language of unity in 1928 Youth Congress.

Aside from using Bahasa Indonesia, most Indonesian people also speak their respective regional languages. Indonesia has the most variety of languages in the world after Papua New Guinea. There are 707 languages that are used as a first language by the Indonesian population (Lewis et al., 2013). According to the Language Development Agency, Ministry of Education and Culture, there were 668 regional languages in Indonesia in 2018. Since the regional languages in the eastern region has not yet been fully identified, this number is expected to continue to increase. The use of regional languages as a first language for the majority of Indonesian population then affects the pronunciation in various regions in Indonesia and affects vocabulary of Bahasa Indonesia.

In addition to being influenced by regional languages, Bahasa Indonesia is also being influenced by foreign languages through trade and religious missions since before the 4th century. These languages include Sanskrit, Tamil and Hindi from India, Chinese, Arabic, Portuguese, Dutch, and English. Their influence can be seen from the large number of Indonesian words originating from foreign languages (Pastika, 2012).

In developing speech technology for Bahasa Indonesia, especially an automatic speech recognition (ASR) and a speech synthesizer, the influences of foreign languages and local languages on the Indonesian language must be considered especially when designing the phoneme set and the vocabulary list and its pronunciation (lexicon). The next section of this paper will explain Indonesian

[§]This work was done while the author was a member of ATR/NICT Spoken Language Communication Research Labs, Japan

phonemes and vocabulary from the linguistic perspective, and how they are used in some research on automatic speech recognition (ASR) systems, and speech synthesizer or text-to-speech (TTS) systems.

2. Bahasa Indonesia

2.1 Grapheme Set

The alphabet used in Indonesian spelling consists of 26 letters; ⟨a⟩ until ⟨z⟩ (PUEBI, 2016). It has a highly phonemic orthography, i.e. almost all graphemes represent one phoneme sound, except for a few sounds represented by digraphs and vice versa, almost all phonemes are represented by either one or two graphemes. Exceptions to this rule are found in three phonemes, which are phonemes /e/ and /ə/ both of which are written with the letter ⟨e⟩ or ⟨E⟩, and phonemes /ʔ/ (glottal stop) which are sometimes written with the letter < k> or not written at all. Other exceptions are in absorption words from foreign languages (Yap, 2010).

2.2 Phoneme Set

According to Soderberg (2008), Bahasa Indonesia has 32 phonemes: vowels (/a/, /e/, /ə/, /i/, /o/, /u/); diphthongs (/ai̯/, /au̯/, /oi̯/); plosives (/b/, /d/, /g/, /k/, /ʔ/, /p/, /t/); affricates (/tʃ/, /dʒ/); nasals (/m/, /n/, /ɲ/, /ŋ/); trill (/r/); fricatives (/f/, /h/, /x/, /s/, /ʃ/, /z/); approximants (/w/, /j/); and lateral approximant (/l/). These phonemes are the standard phonemes used by Indonesians when uttering Indonesian words without considering their allophone.

Sometimes foreign words are also found in daily conversation, especially English words. Because of the foreign language influence, code-switching phenomenon is a common thing found in Bahasa Indonesia. However, most Indonesians pronounce English words using the Indonesian accent, although it varies depending on their English fluency. To cope with the code-switching phenomena, the phoneme set designed for speech technology for Bahasa Indonesia must include the foreign language phoneme analysis. In the CMU (Carnegie Mellon University) pronunciation dictionary there are 39 English phonemes, and some of the phonemes are overlapped with Bahasa Indonesia's phonemes. To determine the phoneme set of English spoken by Indonesian, one of the approaches is by mapping the English phonemes to the most similar phonemes in Bahasa Indonesia as conducted in some speech recognition research for Bahasa Indonesia (Lestari, 2010; Hartanto, 2019). From Lestari (2010), the mappings are as follows (English → Indonesian): /ɑ/→/a/; /æ/→/e/; /ɛ/→/e/; /ʌ/→/ə/; /ɔ/→/o/; /aʊ/→/au̯/; /aɪ/→/ai̯/; /b/→/b/; /tʃ/→/tʃ/; /d/→/d/; /ð/→/d/; /ɝ/ →/e/ /r/; /eɪ/→/e/ /j/; /f/→/f/; /g/→/g/; /h/→/h/; /I/→/i/; /i/→/i/; /dʒ/→/dʒ/; /k/→/k/; /l/→/l/; /m/→/m/; /n/→/n/; /ŋ/→/ŋ/; /oʊ/→/o/; /ɔɪ/→/o/; /p/→/p/; /ɹ/→/r/; /s/→/s/; /ʃ/→/ʃ/; /t/→/t/; /θ/→/t/; /ʊ/→/u/; /u/→/u/; /v/→/f/; /w/→/w/; /j/→/j/; /z/→/z/; /ʒ/→/z/.

2.3 Vocabulary

The development of language reflects the development progress of civilization in a society which can be seen in the development of vocabulary. For the Indonesian language, the development of vocabulary increased rapidly at the end of the 20th century and the beginning of the 21st century which, among other things, was driven by the development of science, technology and the arts. This can be seen from the increase of entries in the official "Great Dictionary of the Indonesian Language" (Indonesian: KBBI) from one edition to the next, and for over the past 20 years, entries in the KBBI increased from 62,000 entries in the first edition (1988) to 91,000 entries in the fourth edition (2008) (Sugono, 2008).

3. Automatic Speech Recognition for Bahasa Indonesia

3.1 Indonesian ASRs and Their Phoneme Set

The first Indonesian ASR was a word-based system covering only a small vocabulary that was developed in 2004 for hearing and speaking impaired people (Sakti et al., 2004). It was performed on acoustic model based on the hidden Markov model with Gaussian mixture model (GMM-HMM). Then, a year later, the initial Indonesian phoneme-based ASR was designed using the cross-language approach, where English is the source language, and Indonesian is the target language (Sakti et al., 2005).

Since there were no agreed standard phoneme set for Indonesian ASR, many researches utilized their own phoneme set. Lestari et al. (2006) utilized 31 phonemes for their GMM-HMM based large-vocabulary ASR. The phoneme set was similar to (Soderberg, 2008) albeit with two differences. Phonemes /e/ and /ə/ were merged because they were spelled as ⟨e⟩. Phoneme /ʔ/ was not used while phoneme /q/ was introduced to accommodate the pronunciation of Arabic loanwords. Sakti et al. (2008a) also developed Indonesian large-vocabulary corpora and a large vocabulary ASR system using 32 phonemes similar to (Soderberg, 2008). They had been carried out for the A-STAR (Asian Speech Translation Advanced Research) consortium; the complete A-STAR system was successfully launched in 2010 (Sakti et al., 2008b; 2013).

Hoesen and Lestari (2014) did not employ the /q/ (since its number was too low and Indonesian tends to pronounce it as /k/) and the diphthongs. Clynes (1997) argued that Bahasa Indonesia (part of the Austronesian language family) did not have diphthongs. The diphthongs were thought of as a combination of a vowel and an approximant. Afterwards, Hoesen et al. (2016a; 2016b) performed experiment on various adaptation methods for the GMM-HMM model. The /q/ and the diphthongs were supplied back; phonemes /e/ and /ə/ were also treated as two separate phonemes. The reintroduction of the diphthongs was to avoid listing monophthongized version of every diphthong.

More recent research started to employ neural-network based acoustic model, the current state-of-the-art method. Hoesen et al. (2015; 2018) utilized shared-hidden-layer fully-connected neural-network (SHL-DNN) which was jointly trained with the high-resource English and the low-resource Indonesian data. The joint training could decrease the recognition errors for the low-resource Indonesian ASR. Although this research used not only Indonesian data, but also English data, phoneme set utilized in these researches were identical with (Hoesen and Lestari, 2014).

3.2 Dictionary and Language Model

Phonetic-based ASR system necessitates the use of phonetic dictionary. Phonetic dictionary translates a word or term into its phonetic sequence(s). Similar to the phoneme set, there is still no standard dictionary (including vocabulary) for Indonesian ASRs. Lestari et al. (2006; 2010) collected unique words from Indonesian news corpus that occurred more than 3 times. Meanwhile, Hoesen et al. (2016a; 2016b) collected unique words from KBBI, various reputable Indonesian news websites, and transcripts from various Indonesian speech corpora.

Some approaches were then performed to generate pronunciation to the words. Hoesen and Lestari (2014) and Hoesen et al. (2018) manually annotated each word in their dictionary. While this approach could yield the best phonetic accuracy, it was impractical for larger dictionary because it needed a considerable amount of time and resources.

As mentioned in Section 2, native Indonesian words tend to have a high degree of consistency between their spelling and pronunciation. Exploiting this tendency, Zahra et al. (2009) and Putri et al. (2019) specifically tried to automatically generated pronunciation for an Indonesian ASR dictionary using a set of rules. Their research yielded 3.2% and 12.71% phone error rates (PERs) respectively. The higher PER from (Putri et al., 2019) was caused by the occurrence of abbreviations and English words and loanwords, which could not be tackled thoroughly by the rules. Hoesen et al. (2019) tried to overcome these problems by employing seq2seq-based neural network to generate the pronunciation. Their cross-validation experiment could achieve 4.15-6.24% PER. To achieve better phone accuracy while requiring less time, automatic pronunciation generation then manual refinement can be performed, such as in (Hoesen et al., 2016a; 2016b).

Other than the dictionary, the majority of current ASRs require a language model (LM). In Indonesian, an LM for informal speech will be dissimilar from an LM for standard formal speech. Informal speech in Indonesian contains some disfluencies, a different set of vocabulary from the formal one, and unusual sentence structures (Hoesen et al., 2016a; 2016b). For example, the suffixes *-i* and *-kan* in formal speech becomes *-in* in informal speech. Complicating matters, Indonesian informal speech is also disparate from the informal text. Indonesian informal texts contain a considerable amount of unpronounceable abbreviations; thus, most informal texts cannot be used for training the spontaneous LM. One method to produce an LM more suitable for informal speech is to adapt a formal speech LM with spontaneous speech transcript (Lestari and Irfani 2015).

4. Automatic Speech Synthesizer for Bahasa Indonesia

Indonesian speech synthesizer (TTS, text-to-speech) has been developed for the past years. The early version of Indonesian TTS systems was developed using di-phone unit concatenation (Arman et al., 2001). Then, Sakti et al. (2008c) developed the first HMM-based speech synthesis for the Indonesian language using only limited resources. The subjective assessment of the Indonesian TTS in terms of both quality and intelligibility aspects had also been conducted online by a web-based listening test system (Sakti et al., 2010).

After that, Jangtjik (2014) developed HMM-based speech synthesizer by adding English lexicon to deal with code-switching, using phone mapping technique i.e mapping English phonemes to their Indonesian equivalent as done by Lestari (2010). Hidayat (2015) improved the naturalness of the TTS by adding prosody model to the system. Improvement in acoustic model was also performed by Fanani (2017), employing DNN to improve overall naturalness. Gisela et. al. (2019) further developed the DNN TTS into an Indonesian-English polyglot TTS to ascertain code-switching problems. Research in Gisela et. al (2019) found that although it improves the intelligibility of the English words, in some cases, the non-polyglot, phone-mapping version of the English words were more easily understood.

4.1 Phoneme Set for Indonesian TTS

The phoneme set used for Indonesian TTS is generally the same, following the set defined by Soderberg (2008), although the vowels are quite simplified compared to the real-life case. The vowels /i, e, o, u/ have allophones /ɪ, ɛ, ɔ, ʊ/ and they might depend on the regional accent of the speaker. Thus, usually in the TTS lexicon, the vowels /i, e, o, u/ represents both themselves and their allophones. The drawback of not including the allophone is that the generated sentences do not sound as natural as native speaker, especially since the allophones are often used in a final closed syllable, for example in the words ‘sindir’ /sindɪr/ and ‘lumpur’ /lumpʊr/, or are used sporadically as in ‘foto’ /foto/ and ‘tokoh’ /tɔkɔh/.

4.2 Vocabulary and Pronunciation Rules for Indonesian TTS

The vocabulary used for Indonesian TTS is quite similar with the ASR lexicon. However, since the TTS models only one speaker, there is no necessity to list multiple pronunciations for a single word. Pronunciations are generally taken from the KBBI. For words that are not provided by the KBBI, letter-to-sound rules are applied, since Indonesian has a quite straightforward grapheme-to-phoneme relationship. Some TTS even rely almost fully on letter-to-sound conversion algorithm, as done by Gisela et. al. (2019), except for words involving the letter ‘e’ and letter pairs ‘ai’, ‘au’, and ‘oi’ that are provided in a small lexicon.

5. Conclusion and Future Works

We have presented a summary of research activities that had been done on the Indonesian language, specifically on the design of the Indonesian phoneme set and two main speech technology applications (ASR and TTS). Nowadays, recent progress in Indonesia's research activities do not focus solely on Indonesian as the official language anymore but start to cover on various Indonesian ethnic languages spoken across the archipelago (i.e., Javanese, Sundanese, Balinese, and Batak languages).

6. Acknowledgements

This report is partially funded by the Ministry of Research and Higher Education of Indonesia under research project

with the title "Intelligent System to Monitor Gadget Usage in Teenagers using Machine Learning Technique". A part of the work on Indonesian ASR/TTS for A-STAR project had been supported by the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT), the Asia Pacific Economic Cooperation Telecommunication and Information (APEC-TEL) Working Group, and the Asia Pacific Telecommunity Standardization Program (APT-ASTAP).

7. Bibliographical References

Arman, A. (2001). Prosody model for Indonesian text to speech system. Asia Pacific Conference on Communication, Tokyo.

Clynes, A. (1997). On the Proto-Austronesian "Diphthongs". Oceanic Linguistics, vol. 36(2), pages 347-362.

Hartanto R., Lestari D.P. (2019). Rule-based Approach for English-Indonesian Code-switching Acoustic Model. International Conference on Data and Software Engineering 2019.

Hasan, Alwi. Dendy, Sugono. Anton, Moeliono (1999). Telaah Bahasa dan Sastra. Yayasan Obor Indonesia, page 260.

Hoesen, D. and Lestari D.P. (2014). A Prototype of Indonesian Dictation Component for Typing and Formatting Document Using a Word Processor Software, Proceedings of International Conference on Electrical Engineering and Computer Science, pages 17-21.

Hoesen, D., Satriawan, C.H., Lestari, D.P., and Khodra, M.L. (2016a). Towards Robust Indonesian Speech Recognition with Spontaneous-Speech Adapted Acoustic Models. Procedia Computer Science, vol. 81, pages 167-173.

Hoesen, D., Lestari, D.P., and Khodra, M.L. (2016b). Adaptation of Acoustic Model for Indonesian Using Varying Ratios of Spontaneous Speech Data. Proceedings of the 2016 O-COCOSDA, pages 39-44.

Hoesen D., Price R., Lestari D.P., Shinoda K. (2015) A DNN-based ASR system for the Indonesian language. Proceedings of ASJ Autumn Meeting, pages 5-6.

Hoesen D., Lestari D.P., and Widyantoro D.H. (2018) Shared-hidden-layer deep neural network for under-resourced language. Telkomnika, vol. 16(3), pages 1226-1238.

Hoesen D., Putri, F.Y., and Lestari D.P. (2019) Automatic Pronunciation Generator for Indonesian Speech Recognition System Based on Sequence-to-Sequence Model. Proceedings of the 2019 O-COCOSDA.

Lestari, D.P., Iwano K., and Furui S. (2006). A large vocabulary continuous speech recognition system for Indonesian language. Proceedings of the 15th Indonesian Scientific Conference in Japan.

Lestari, D.P., Furui S. (2010). Adaptation to Pronunciation Variations in Indonesian Spoken Query-Based Information Retrieval, IEICE Transactions on Information and Systems, 93(9), pages 2388-2396.

Lestari, D.P. and Irfani A. (2015). Acoustic and language models adaptation for Indonesian spontaneous speech recognition. Proceedings of the 2nd International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA).

Lewis, M. Paul, Gary F. Simons, and Charles D. Fennig (2013). Ethnologue: Languages of the World, Seventeenth edition. Dallas, Texas: SIL International.

Pastika, I. Wayan (2012). Pengaruh Bahasa Asing terhadap Bahasa Indonesia dan Bahasa Daerah: Peluang atau Ancaman?. Jurnal Kajian Bali, Vol. 02, No. 02, ISSN: 2580-0698.

Panitia Pengembang Pedoman Bahasa Indonesia (2016). Kementerian Pendidikan dan Kebudayaan. Pedoman Umum Ejaan Bahasa Indonesia. 4th ed.

Putri, F.Y., Hoesen D., and Lestari D.P. (2019). Rule-Based Pronunciation Models to Handle OOV Words for Indonesian Automatic Speech Recognition System. Proceedings of the 6th International Conference on Science in Information Technology (ICSITech).

Yap M.J. et al., (2010). The Malay lexicon project: A database of lexical statistics for 9,592 words., Behavior Research Methods, 42(4), pages 992-1003.

Sakti, S., Hutagaol, P., Arman, A.A., and Nakamura, S. (2004). Indonesian speech recognition for hearing and speaking-impaired people. International Conference on Spoken Language Processing. pages 1037–1040.

Sakti, S., Markov, K., and Nakamura, S. (2005). Rapid Development of Initial Indonesian Phoneme-based Speech Recognition Using the Cross-Language Approach. O-COCOSDA.

Sakti, S., Kelana, E., Riza, H., Sakai, S., Markov, K., Nakamura, S. (2008a). Recent Progress in Developing Indonesian Large-Vocabulary Corpora and LVCSR System. MALINDO. pages 40-45.

Sakti, S., Kelana, E., Riza, H., Sakai, S., Markov, K., Nakamura, S. (2008b). Development of Indonesian Large Vocabulary Continuous Speech Recognition System within A-STAR Project., IJCNLP Workshop on TCAST, pages 19-24.

Sakti, S., Maia, R., Sakai, S., Shimizu, T., Nakamura, S. (2008c). Development of HMM-based Indonesian Speech Synthesis. O-COCOSDA. pages 215-220.

Sakti, S., Sakai, S., Isotani, R., Kawai, H., Nakamura, S. (2010). Quality and Intelligibility Assessment of Indonesian HMM-Based Speech Synthesis System. MALINDO, pages 51-57.

Sakti, S., Paul, M., Finch, A., Sakai, S., Vu, T.-T., Kimura, N., Hori, C., Sumita, E., Nakamura, S., Park, J., Wutiwiwatchai, C., Xu, B., Riza, H., Arora, K., Luong, C.-M., Li, H. (2013). A-STAR: Toward Translating Asian Spoken Languages. Special issue on Speech-to-Speech Translation, Computer Speech and Language Journal (Elsevier), vol. 27, Issue 2, pages 509-527.

Soderberg C.D. & Olson K.S., (2008). Indonesian, Journal of the International Phonetic Association, 38(2), pages 209-213.

Sugono, Dendy., et al., (2008). Kamus Pusat Bahasa. Jakarta. ISBN 978-979-689-779-1

Zahra, A., Baskoro, S., and Adriani M. (2009). Building a pronunciation dictionary for Indonesian speech recognition system. Proceedings of Workshop on Technologies and Corpora for Asia-Pacific Speech Translation (TCAST).

Proceedings, pages 210–214
Paris, UNESCO Headquarters, 5-6 December, 2019.

Analysis of the Ethiopic Twitter Dataset for Abusive Speech in Amharic

Seid Muhie Yimam[1], **Abinew Ali Ayele**[12], **Chris Biemann**[1],
Language Technology Group, Department of Informatics, MIN Faculty
Universität Hamburg, Germany[1],
Faculty of Computing, Bahir Dar Institute of Technology
Bahir Dar University, Ethiopia[2]
{yimam, ayele, biemann}@informatik.uni-hamburg.de

Abstract

In this paper, we present an analysis of the first Ethiopic Twitter Dataset for the Amharic language targeted for recognizing abusive speech. The dataset has been collected since 2014 that is written in Fidel script. Since several languages can be written using the Fidel script, we have used the existing Amharic, Tigrinya and Ge'ez corpora to retain only the Amharic tweets. We have analyzed the tweets for abusive speech content with the following targets: Analyze the distribution and tendency of abusive speech content over time and compare the abusive speech content between a Twitter and general reference Amharic corpus.

Keywords: abusive speech, hate speech, offensive speech, less-resourced language, Amharic tweet

ረቂቅ

በዚህ ጽሑፍ ውስጥ የጥላቻ ንግግሮችን ለመለየት ለሚደረጉ ጥናቶች የሚያገለግል የአማርኛ ቋንቋ የመጀመሪያ የትዊተር የውሀብ ስብስብ ዳሰሳ ጥናት አቅርበናል። የኢትዮጲክ የትዊተር ማህበራዊ ትስስር የውሀብ ስብስብ ከ2014 እ.ኤ.አ ጀምሮ ተሰብስቧል። በኢትዮጲክ ፊደል (ግዕዝ ፊደል) የተጻፉ ትዊቶች ብቻ ተለይተው በአንድ የመረጃ ቋት ውስጥ ተቀምጠዋል። የግዕዝ ፊደልን በመጠቀም የሚፃፉ በርካታ ሌሎች ቋንቋዎችም ስላሉ፤ የአማርኛ ትዊቶችን ብቻ ለመለየት አሁን ላይ የሚገኙ የአማርኛ ፣ የትግርኛ እና የግዕዝ የፅሁፍ ስብስቦችን ተጠቅመናል። የጥላቻ ንግግሮችን ይዘት በተመለከተ ከትዊተር የተገኘውን ፅሁፍ የዳሰሰነው በሚከተሉት አቅጣጫዎች ነው:- 1) የጥላቻ ንግግሮችን ይዘት፤ ስርጭት እና ዝንባሌ ከጊዜ ሂደት ጋር መተንተን፤ 2) በትዊተር የማህበራዊ ትስስር ጽሑፍ እና በአጠቃላይ የአማርኛ ማጣቀሻ የመረጃ ስብስብ መካከል ያለውን የጥላቻ ንግግሮች ይዘት ማነፃፀር።

1. Introduction

The emergence of social media creates seamless communication between people and hugely increases the level of information sharing. In the Ethiopian case, people use social media as a primary source of information, and they tend to believe everything from these sources. Recently, we have witnessed a large level of chaos in Ethiopia due to misinformation and abusive language dissemination using social media. The hate speech and fake news dissemination already affected the lives of millions, schools and universities recently closed, business activities heavily hampered due to closure of main roads in the country, the movement of citizens has been seriously hindered, and millions are displaced while hundreds have died (Kiruga, 2019).

It is now a global trend to fight the dissemination of false news and abusive language. Some of the nations have already created regulations that should be compliant with freedom of speech[1] (Levush, 2019).

At the beginning of 2019, the Ethiopian government has drafted legislation[2] against hate speech and hold a series of discussions with different stakeholders, where it is expected to be a law once approved by the parliament before the end of the year.

In this paper, the primary focus is to briefly analyze the Ethiopic Twitter Dataset (ETD) towards abusive speech for Amharic. We hope that this paper, in general, serves as a basis for future research concerning social media contents and, in particular, to study the abusive speech usage and trends in social media for the Amharic language. It further opens a dialogue between technology practitioners, law enforcement parties, and citizens as well on how to deal, regulate and counter attack abusive speech using social media[3].

2. Motivation of the Study

The emergence of social media, particularly Facebook and Twitter facilitate the way people communicate in their day-to-day activities. It makes the communication and sharing of information much faster and easier. It brings a friend closer than ever, which otherwise not possible to maintain such links. In the case of Ethiopian social media communication, it is believed that the connection between the larger population of the Diaspora and the friends at home is getting much easier. Furthermore, it has facilitated the transfer of knowledge and technology much simpler and more affordable.

Despite such huge positive influences, social media is bringing its negative consequences to the Ethiopian population than other developing countries (Sibhat, 2018). Social

[1] https://www.poynter.org/ifcn/anti-misinformation-actions/

[2] https://bit.ly/2KDSVDx

[3] This paper tries to highlight the coverage of abusive languages on social media content based on a list of keywords collected form limited audiences. We do not yet conduct a proper abusive language analysis and can not also declare a given word, phrase, or sentence as an abusive or not. Moreover, topics discussed are not based on a specific law from the Ethiopian constitution, rather they are based on a general and technological notion that is adopted in the global arena of hate and offensive speech research.

media makes the dissemination of rumors, false information, and hate speech much faster, as a larger portion of the population is already using smartphones for their daily communications.

The article by Dibaba (2019) pointed out that the dissemination of hate speech is endangering the democratic rights, jeopardize the long-standing social fabrics and ultimately create political and socio-physiological havoc destabilizing the country. The definition of abusive texts in this paper is confined to the definition of the new draft regulation that is proposed by the Ethiopian government this year.

2.1. The New Ethiopian Draft Regulation about Hate Speech

The socio-political crisis that existed since 2016 in Ethiopia, has caused devastating ethnic and sometimes religious-based conflicts. Many people died, displaced from their villages, private and government buildings were also destroyed. The role of hate speech spanning through social media in aggravating these devastating mass conflicts was paramount. It has been noted that hate speech, in the current polarized Ethiopian politics escalates the danger of ethnic and sometimes religious-based mass conflicts by inciting the public (Sibhat, 2018).

In April 2019, the attorney general of the Federal Democratic Republic of Ethiopia has prepared a draft law[4] to tackle hate speech and fake news. In this 5 page draft, which is prepared in Amharic, it describes what defines hate speech and fake news in more general terms. Particularly, it defines *hate speech* as a speech that targets an individual, group or community based on religion, race or color, gender or physical appearance, immigration or origin, and language that intentionally depicts the target as evil, demeans, threatens, discriminates, or otherwise evoke violence.

In this regard, hate speech is targeting a certain ethnic or a specific political group and religion that jeopardizes the exercise of human and democratic rights in the country. Moreover, hate speech threatens the peaceful social life, the long-lasting unity of people and even may lead to a massive massacre between ethnic as well as religious groups if not managed by regulation. Therefore the need for a regulation to govern hate speech is very critical and timely (Dibaba, 2019).

However, the draft is criticized as being poorly drafted with profound implications for human rights in general and freedom of expression as well as the right to privacy in particular. The draft is also blamed for confusing social media with the conventional media[5]. It also fails to impose clear criminal responsibility on hatred social media users and many other vague and confusing even unseen scenarios that should seriously be considered (Abraha, 2019).

3. Dataset Collection

3.1. General Reference Corpora

While our main purpose is to analyze the content of the ETD for abusive languages in Amharic, we also collect and analyze general reference corpora (GRC) mainly 1) used to train language models for language identification tasks, and 2) to examine the distribution of the selected keywords for abusive language. Even though there are more than 10 Ethiopian languages that use the Ethiopic script (the Fidel) for their writing system, we have obtained a textual dataset only for three languages, namely Amharic (GRC-AM), Tigrinya (GRC-TI), and Ge'ez (GRC-GE). The size and description of these corpora are presented in Section 3.3.

3.2. Twitter Dataset

The Ethiopic Twitter Dataset for Amharic (ETD-AM), which is the main focus of analysis in this paper, is collected from mid-August 2014 and continues collecting the tweets written in Fidel script every day. We have collected specifically texts written with Fidel script. Our program runs every day and fetches the tweet, date, time, user location, tweet ID. Until now, around three million tweets have been collected from 154,477 users.

3.3. Language Identification and Separation

Since the Fidel script used as writing system for various **Ethiopian** and **Eritrean** languages, such as **Argobba**, **Awngi**, **Blin**, **Chaha**, **Dizin**, **Harari**, **Inor**, **Silt'e**, **Tigre**, **Tigrinya** and **Xamtanga**[6], we have developed a language identification and separation component. There is no publicly available tool to detect and identify texts written in Fidel script into their respective language families (Semitic languages). For the three Ethiopic languages, namely Amharic, Tigrinya, and Ge'ez, there are corpora of sufficient size that can be used to train a model for language detection.

Amharic and Tigrinya are currently used both in academic and daily information propagation (mainly traditional news outlet and social media texts) while Ge'ez is mainly used in the production and dissemination of religious texts by the Ethiopian Orthodox Church (Molla, 2018). We suppose that the ETD we have collected is a mixture of mainly these three languages. To identify the languages of each tweet, we build a language model based on the work of Cavnar and Trenkle (1994), which uses N-gram frequency statistics. For Amharic, we have texts from three sources, 1) web-corpus texts that we have collected at Universität Hamburg using a focused crawler, 2) from the Opus repository[7] (Tiedemann, 2012) where they have more than 300 parallel corpus text, and 3) from the Amharic web corpus (Suchomel and Rychlý, 2016a). For Tigrinya, we use texts from the Opus repository and the Tigrinya web corpus (Suchomel and Rychlý, 2016b). For Ge'ez, we have manually crawled religious books from the Scripture Tools for Every Person (STEP)[8] and from the Lexical Data Repository of the Ge'ez Frontier Foundation[9]. Those tweets other than this stated three languages are categorized as other and trimmed out from our analysis since we do not find available datasets to

[4] https://bit.ly/2KDSVDx

[5] https://theowp.org/ethiopias-drafted-legislation-against-hate-speech-threatens-journalistic-freedoms/

[6] https://www.omniglot.com/writing/ethiopic.htm

[7] http://opus.nlpl.eu/

[8] https://www.stepbible.org/version.jsp?version=Geez

[9] https://github.com/geezorg/data

build the respective language identification model. Table 1 shows the statistics of the three corpora (upper half) and the distribution and statistics of tweets identified into the three languages (lower part) while Table 2 displays the top 5 frequent n-grams.

Language	Tokens	Types
GRC-AM	46,353,602	1,363,192
GRC-TI	8,512,177	339,189
GRC-GE	316,740	42,721
ETD-AM	26,277,724	1,097,986
ETD-TI	3,152,168	309,851
ETD-GE	385,336	52,114
ETD-Other	195,326	19,777

Table 1: The number of tokens and types (unique occurrences of tokens) in the ETD and GRC dataset. The suffix *-AM*, *-TI*, and *-GE* stands for *Amharic*, *Tigrinya*, and *Ge'ez* respectively. In the Ethiopic Twitter Dataset, when the text written in the Fidel script can not be identified as either *Amharic*, *Tigrinya*, or *Geez*, it is placed in a separate group as *Other*.

4. Abusive Language in ETD-AM

In this section, we will analyze the nature and distribution of abusive texts in Amharic using the ETD based on keywords collected from 5 native speakers. The tweets we use for the analysis are only the Amharic tweets that are identified and filtered by the language model.

In the following sub-section, we will analyze particularly the emergence and proliferation of abusive speech on Twitter. All charts show normalized frequencies in the unit of parts per million (ppm).

4.1. Keywords for Abusive speech

In this paper, we adopt the definition of hate and offensive speech based on the work of Davidson et al. (2017). The distinction between hate and offensive speech is always blurry, and we believe that it also depends on the languages, situations or context, and times of the events. We define keywords as hate speech if it fits the definition of the current draft legislation. Otherwise, we categorize the keywords as offensive speech.

We have collected 99 hate speech and 48 offensive speech keywords for the Amharic language from different participants (native speakers)[10]. The participants have collected the keywords from Facebook posts and comments, Twitter tweets and re-tweets, and Youtube comments from popular pages.

4.2. Analysis of Abusive Speech in Amharic

Based on the keywords we have collected, we have analyzed the ETD-AM from different aspects. Since the dataset has been collected for 5 years, we first analyze how the keywords are distributed in the dataset. The ETD collected in 2015 were not correctly stored in our database due to an encoding issue. Hence we do not analyze the dataset for this year.

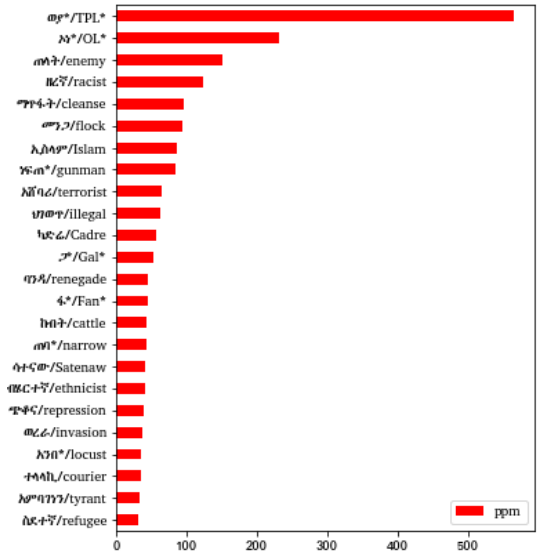


Figure 1: Distribution of hate speech keywords

Figure 1 shows the frequencies of hate speech keywords while Figure 2 shows the frequencies of offensive keywords in the Amharic Twitter dataset. From these figures, we can see that the frequencies of hate speech keywords are very large compared to their offensive counterparts.

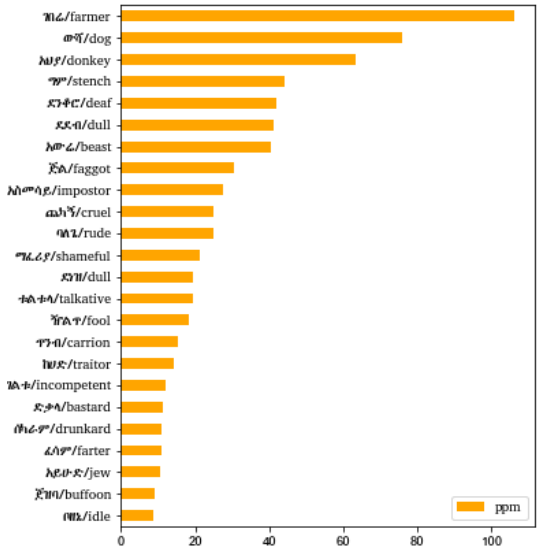


Figure 2: Distribution of offensive speech keywords

From Figure 3, it can be seen that the number of tweets is increasing over time. The same holds for the number of people using Twitter social media are also increasing continuously.

An interesting analysis is observed when we compare the distribution of hate and offensive speech keywords in the ETD-AM and GRC-AM. Even if there are quite a large

[10] We select participants who are actively engaging in social media and who are from different fields of study (Political science, Journalism, Engineering, Business administration, and Computational linguistics).

Uni-grams		Bi-grams		Tri-Grams	
Word	Freq.	Phrase	Freq.	Phrase	Freq.
ነው/is	346,965	አዲስ አበባ/Addis Ababa	13,538	ዶ/ር አብይ አህመድ/Dr. Abiy Ahmed	3,778
ላይ/on	138,466	አብይ አህመድ/Abiy Ahmed	10,954	ላይክ እና ሼር/like and share	3,718
እና/and	125,040	ብቻ ነው/and only	9,372	እርሶም ትኩስ መረጃዎችን/you too hot-news	2,066
ግን/but	60,580	በአዲስ አበባ/by Addis Ababa	8,262	ጠ ሚ አብይ/PM Abiy	1,963
ሰው/man	56,502	የአዲስ አበባ/of Addis Ababa	8,185	እንኳን ደስ አለዎት/Congratulations	1,917

Table 2: The most five frequent Ngrams from Amharic tweets

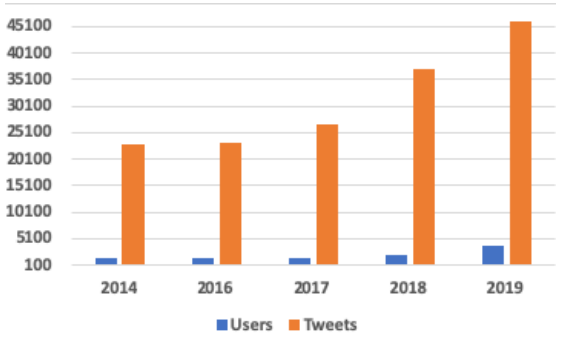


Figure 3: Number of users and Amharic tweets in the ETD per year for the last five years

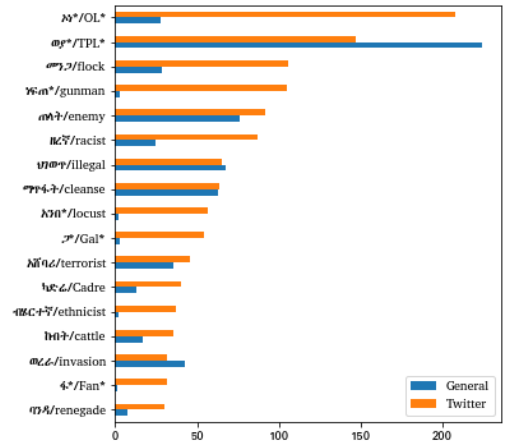


Figure 4: Comparison of hate speech keywords (y-axis) between the GRC-AM and the ETD-AM, based on their respective ppm (x-axis).

number of texts in the general domain, particularly abusive keywords have occurred more often in the ETD-AM than the GRC-AM (See Figure 4 and 5). Keywords that are particularly used conventional news portals such as organization names (example TPLF) are more dominant in the GRC-AM dataset than in the ETD-AM dataset. Whereas, if the organization name is labeled as abusive by the mainstream media (example OLF), the term appears more in the ETD than in the GRC-AM dataset.

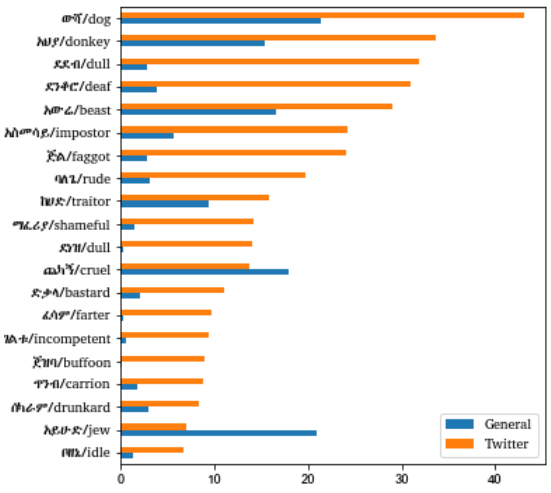


Figure 5: Comparison of offensive speech keywords (y-axis) between the GRC-AM and the ETD-AM, based on their respective ppm (x-axis).

5. Conclusion

In this paper, we report the distribution of abusive speech for the Amharic language based on the Ethiopic Twitter Dataset. We have collected around 144 abusive speech keywords from 5 native speakers and categorize them into hate and offensive speech. We then analyze how abusive speech develop over the last five years. In general, the total amount of Amharic Tweets, as well as the number of tweets containing abusive keywords, are increasing over time. The dataset will be used to build automatic abusive language detection systems for Amharic.

6. Bibliographical References

Abraha, H. H. (2019). The problems with Ethiopia's proposed hate speech and misinformation law. *The London School of Economics and Political Science, June 4, 2019*. Url: https://blogs.lse.ac.uk/medialse/2019/06/04/the-problems-with-ethiopias-proposed-hate-speech-and-misinformation-law/.

Cavnar, W. B. and Trenkle, J. M. (1994). N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, Las Vegas, Nevada, USA.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the 11th*

International AAAI Conference on Web and Social Media, ICWSM '17, pages 512–515, Montreal, Canada.

Dibaba, S. (2019). Hate speech and freedom of expression in Ethiopia. *The Ethiopian Herald, May 8/2019*. Url: https://www.press.et/english/?p=5520#.

Kiruga, M. (2019). Ethiopia struggles with online hate ahead of telecoms opening. *The African Report*. Url:https://www.theafricareport.com/19569/ethiopia-struggles-with-online-hate-ahead-of-telecoms-opening/.

Levush, R. (2019). Limits on Freedom of Expression. *Law Library of Congress*. Url:https://www.loc.gov/law/help/freedom-expression/limits-expression.pdf.

Molla, E. D. (2018). An analysis of Ge'ez language heritage potential: traditional church schools and the practices of Ethiopian Orthodox Tewahido Churches. *Pharos Journal of Theology ISSN 2414-3324*. Url:https://www.pharosjot.com/uploads/7/1/6/3/7163688/article_21_vol_99_2018_-__ethiopia.pdf.

Sibhat, H. N. (2018). Spreading Hatred A study of Facebook in Ethiopia. *Global Media Review (GMR), Vol-1, Issue-2*.

Suchomel, V. and Rychlý, P. (2016a). Amharic Web Corpus. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Suchomel, V. and Rychlý, P. (2016b). Tigrinya Web Corpus. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)* , pages 2214 – 2218, Istanbul, Turkey.

Proceedings, pages 215–218
Paris, UNESCO Headquarters, 5-6 December, 2019.

The IRCAM Realizations for the Amazigh Preservation and Revitalization in Morocco

Fadoua ATAA ALLAH, Aicha BOUHJAR
CEISIC, DCOM
The Royal Institute of Amazigh Culture
Allal El Fassi Avenue, Madinat Al Irfane, Rabat, Morocco
{ataaallah, bouhjar}@ircam.ma

Abstract

The computerization of languages is a strategic issue, closely linked to the rise of information and communication technologies. The globalization of exchanges and the dematerialization of information due to the digital revolution have changed the status of several languages in the world and have influenced their cultural, industrial and economic stakes. Aware of this, the Royal Institute of Amazigh Culture has put in place a vision for the Amazigh digitalization. This vision consists in computerizing the language, from a simple display on screens to the development of specialized tools. This paper presents the Institute's achievements for Amazigh to live in the "information society".

Keywords: Amazigh, Less Resourced Language, Revitalization

ⵜⴰⴳⴹⵡⵉⵜ

ⵜⴰⵙⵏⵎⴰⵍⴰⵢⵜ ⵏ ⵜⵓⵜⵍⴰⵢⵜ ⵜⴳⴰ ⵢⴰⵏ ⵓⵙⵇⵙⵉ ⴰⵙⵜⵕⴰⵜⵉⵊⵉ ⵉⵟⵟⴼⵏ ⵙ ⵡⴰⵍⴰⵢ ⵏ ⵜⵉⵜⵉⴽⵏⵓⵍⵓⵊⵉⵜⵉⵏ ⵏ ⵓⵙⵏⵖⵎⵙ ⴷ ⵓⵎⵢⴰⵡⴰⴹ. ⵜⴰⵙⵎⴰⴹⴰⵍⵜ ⵏ ⵉⵎⵙⴽⴰⵍⵏ ⴷ ⴳⴰⵔ ⴰⵎⴰⵜⵜⵉⵡ ⵏ ⵓⵙⵏⵖⵎⵉⵙ ⵉⵣⵍⵉⵏ ⵙ ⵜⵏⴽⵔⴰ ⵜⴰⵎⴰⵟⵟⵓⵏⵜ, ⵙⴰⵜⵜⵉⵏ ⴰⴷⴷⴰⴷ ⵏ ⵡⴰⵟⵟⴰⵚ ⵏ ⵜⵓⵜⵍⴰⵢⵉⵏ ⴳ ⵓⵎⴰⴹⴰⵍ, ⵢⵉⵍⵉ ⴷⴰⵔⵙⵏ ⵢⵉⴹⵉⵚ ⴳⴼ ⵉⵎⵙⴰⵔⴰⵙⵉⵏ ⵉⴷⵍⵙⴰⵏⵏ ⵉⵎⴳⵓⵔⴰⵏⵏ ⴷ ⵉⴷⴰⵎⵙⴰⵏⵏ. ⵉⴷⴷⵖ ⵉⴼⴰⴼⴰ ⴳⴼ ⵓⵢⴰⴷ, ⵉⵙⴱⴷⴷⴰ ⵓⵙⵉⵏⴰⴳ ⴰⴳⵍⴷⴰⵏ ⵏ ⵜⵓⵙⵙⵏⴰ ⵜⴰⵎⴰⵣⵉⵖⵜ ⵢⴰⵜ ⵜⴰⵏⵏⴰⵢⵜ ⵜⴰⵙⵏⵎⴰⵍⴰⵢⵜ ⵏ ⵜⵎⴰⵣⵉⵖⵜ. ⵜⴰⵏⵏⴰⵢⵜ ⴰⴷ ⵜⵣⵍⵉ ⵙ ⵓⵏⵙⵎⴰⵍⴰ ⵏ ⵜⵓⵜⵍⴰⵢⵜ :ⵣⴳ ⵓⵙⵎⴰⵍ ⵏ ⵓⵎⵉⵣⴰⵕ ⴰⵔ ⴰⵙⴱⵓⵖⵍⵓ ⵏ ⵉⵎⴰⵙⵙⵏ ⵉⵙⵜⵉⵏ. ⴰⵔⵔⴰ ⴰⴷ ⵉⵙⵎⵏⵉⴷ ⵉⵙⵓⴼⵖⵏ ⵏ ⵓⵙⵉⵏⴰⴳ ⵅⵎⴰ ⴰⴷ ⵜⵣⴹⴰⵕ ⵜⵓⵜⵍⴰⵢⵜ ⵜⴰⵎⴰⵣⵉⵖⵜ ⴰⴷ ⵜⵜⴷⵔ ⴳ ⵡⴰⵎⵓⵏ ⵏ ⵓⵙⵏⵖⵎⵙ.

1. Introduction

Throughout human history, minority languages are replaced by politically, economically, or socio-culturally dominant ones. This is due to many factors that, every day, endanger certain languages and threaten their transmission. These factors can be structured in three main reasons:

- Language extinction caused by the decimation of its community.
- Eradication of minority languages by certain linguistic policies.
- Renunciation to transmit a language to future generations, so that they are not disadvantaged by being socially excluded.

Some Moroccan languages have also suffered from these factors. According to the UNESCO Project: Atlas of the World's Languages in Danger, seven Moroccan languages are in danger, two of them have completely disappeared. These are the Judeo Berber and the Tamazight of Aït Rouadi[1].

Unfortunately, current technological changes may, if no remedy is taken, dramatically increase this linguistic impoverishment. In this context, the Royal Institute of Amazigh Culture (RIAC; IRCAM in French) is committed to the preservation and the revitalization of the Amazigh language.

This paper aims to present the IRCAM's achievements. Thus, the remaining of this paper is organized as follow: Section 2 introduces the Amazigh sociolinguistic context. Section 3 presents the Royal Institute of Amazigh Culture. Section 4 gives a brief overview about the IRCAM's realizations, especially those related to language planning and language technology. Finally, Section 5 draws conclusion and exposes some future perspectives.

2. Sociolinguistic context

Amazigh language (known as Berber language in the literature) stretches, in a geographically discontinuous manner, over a vast area of North Africa, from the Canary Islands to the Siwa Oasis in Egypt in the north and from the Mediterranean coast to Niger, Mali and Bukina Fasso in the south.

Figure 1: Amazigh language distribution over North Africa[2]

---

[1] http://www.unesco.org/languages-atlas/fr/atlasmap.html

[2] https://www.languagesoftheworld.info/language-contact-2/many-arabic-varieties-traces-roman-heritage-moroccan-arabic.html

Historically, the Amazigh language has been autochthonous and is a member in the Hamito-Semitic or "Afro-Asiatic" family. On the linguistic side, the language is characterized by the proliferation of dialects due to historical, geographical and sociolinguistic factors. This language has its specific writing system, called tifinagh, but it was neglected except in the Saharian region where it was transmitted, mainly by women, without discontinuity, from antiquity up till now. In the other areas, Amazigh was exclusively spoken and reserved for familial and informal domains. The old tifinagh script is found engraved in stones and tombs in some historical sites attested from 40 centuries. Its writing form has continued to change from the traditional Tuareg writing to the neo-tifinagh in the end of the sixties.

Figure 2: Tinzouline Inscriptions (Zagora, Morocco)

In Morocco, one may distinguish three major dialects on the basis of their geographical situation: Tamazight of the North (known as Tarifit or Rifian variety by linguists); Tamazight of Central Morocco and South-East, and Tachelhiyt in the South-West and the High Atlas. According the last demolinguistic data (2004), the Amazigh language is spoken by some 30% of the Moroccan population (around 10 million inhabitants). Politically, the Amazigh language has enjoyed its status at different levels, depending on the country where this language exists. In Morocco, the status of Amazigh has achieved an advanced level since 2011: its officialization, alongside Arabic, in the new Constitution. But the outline of the Amazigh language preservation and promotion is mentioned since 2001 in the text creating and organizing the Royal Institute of Amazigh Culture (Dahir or Royal Decree of 17th October 2001[3]).

3. The Royal Institute of Amazigh Culture

This institution came as an answer to the civil society request since the very beginning of the 60 (after independence) which asked the recognition of the Amazigh language and culture by the State and its institutionalization in education, the media, culture and public administration. IRCAM is under royal tutelage and is dedicated to the promotion of the Amazigh language and culture.

It has the following missions:

- Advisory mission to the Royal Cabinet on measures meant to promote Amazigh.
- Partnership mission with the concerned institutions, in particular the Ministry of National Education, Ministry of Information and Communication, Ministry of Culture, Ministry of Justice, Ministry of the Interior and Ministry of Public Service.
- Academic missions: collection and transcription of various Amazigh cultural expressions with an eye to safeguarding, protecting and disseminating these expressions; studies and research on Amazigh culture; promotion of artistic creation; codifying the Amazigh graphic system for teaching ends, production of didactic tools, elaboration of general and specialized lexicons, elaboration of pedagogical plans of action; cooperation with universities in organizing research and Amazigh language and culture development centers, training trainers; development of methods meant to strengthen and encourage the place of Amazigh in communication and information spaces; cooperation with cultural and scientific institutions at the national and international levels.

During these two last decades, changes occur in the sociolinguistic field and are specifically related to the Amazigh language. What are the observed changes? How was managed the Amazigh language planning process in Morocco? Which were the priorities and what principles guided the decision making? These questions, among others, had to be answered by IRCAM in order to contribute to the new language policy of the country.

Indeed, the motive 6 of the founding text of IRCAM explains that it has "to deepen the linguistic policy defined by the National Charter for Education and Training (CNEF), which stipulates the introduction of Amazigh in the educational system". Indeed, Amazigh, in the National Charter, was considered as an "opening", a language facilitating the learning of the Arabic language (COSEF, 2000: lever 9 paragraphs 115 and 116). With the creation of IRCAM, the ultimate goal is to introduce the Amazigh language, not as a facilitator but as a language taught and a "national heritage" of all Moroccan pupils and students. The "codification of the Amazigh script will facilitate its teaching, learning and diffusion, guarantee equal opportunities for all the children of our nation in acquiring knowledge, and consolidate national unity" (Dahir, 2001). Since then, the Amazigh has become, alongside Arabic, the official language included in the new Constitution.

As it can be observed, within a decade, the status of the Amazigh gradually has changed: from "opening" (CNEF, 1999) to official language (Constitution, 2011: Art. 5[4]) through a process of patrimonialization and language planning (Dahir, 2001).

IRCAM had to translate the general orientations stipulated in the Dahir into concrete action to enable the introduction of the Amazigh language and culture in the public sphere but primarily in the educational and training system. Therefore, issues related to which Amazigh to teach and with which script have arisen as an urgent and pressing priority. In other words, IRCAM had to deal with the "corpus" of the Amazigh language in Morocco.

[3]https://www.ircam.ma/?q=fr/node/4668

[4]http://www.sgg.gov.ma/Portals/0/constitution/constitution_2011_Fr.pdf

4. IRCAM Realizations

Amazigh language has rich oral corpus with stories, songs and histories that was threatened. To preserve this heritage, it is important to revitalize the language (Boukous, 2012) and ensure it transmission. In this aim, IRCAM piloted and drew up a strategy that has been translated into actions through steps involving the progressive planning of Amazigh, its insertion into the educational system and its integration into the digital world. The following points present some of these actions.

4.1 Informatization Roadmap

To ensure the integration of Amazigh language in the information technology sphere and to guarantee its sustainability, IRCAM has elaborated a roadmap. This roadmap is organized in basic processes, designed to take place in three phases: short, medium and long terms.

These processes are represented by a chain starting from elementary processing, passing by the language resources constitution, and going towards generic applications. This chain ranges from adapting and improving tools based on new technologies to the development of applications (Ataa Allah and Boulaknadel, 2014a).

The main bricks of the Amazigh informatization roadmap introduced below are graphic encoding, language tools and resources, in addition to language learning materials.

4.2 Graphic Encoding

Before the creation of IRCAM, the Amazigh language has been writing, in Morocco, in Latin alphabet or in Arabic script. In order to facilitate the decision making process regarding the script choice, IRCAM produced a technical report analyzing the two different scripts in use in addition to the Amazigh ancestral writing system (tifinagh script).

After the adoption of tifinagh as the official script for writing the Amazigh language, in Morocco, on February 10th2003, the Unicode homologation of this script has become a necessity (Andries, 2008). This encoding enabled Amazigh language to have a prominent position in the digital world both at the national and international levels. Beside tifinagh Unicode standardization, Amazigh was integrated also in the international standard prescription keyboards ISO/IEC 9995, to facilitate its keyboarding.

To ensure the conversion of all the Amazigh writing into a standard and a unique form, which is Unicode based tifinagh, a transliterator was developed (Ataa Allah and Boulaknadel, 2011). Moreover, to help visually impaired persons to learn Amazigh language, IRCAM proposed a Braille-based tifinagh writing system and built a Tifinagh-Braille convertor (Yakoubi et al., 2016).

4.3 Language Tools and Resources

Since the colonial period, many studies have been undertaken. They have contributed to the collection of the Amazigh lexicon and oral tradition, and have focused on linguistic features. However, most of these studies covered local dialects only.

The inclusion of the Amazigh language in public life, in Morocco, has resulted in the extension of its areas of use, which has generated the need for terminology, and a reference grammar.

Thus, IRCAM has produced sectorial lexicons, including grammatical, media and administrative terminologies (Ameur et al., 2009ab; Ameur et al., 2013; Ameur et al., 2015), in addition to an Amazigh language general dictionary (Ameur et al., 2017). Moreover, it has published a standard morphosyntactic reference books where the spelling rules were set (Boukhris et al., 2008, Laabdelaoui et al., 2012).

In parallel, IRCAM has undertook projects to build terminology and lexical databases. The terminology database contains terminological entries related to grammatical, media and administrative themes[5]. This database was used also in the production of the 'LEXAM' mobile application[6] (Frain et al., 2014). Whereas, the lexical database includes the usual lexicon belonging to the Amazigh variants[7] (Ataa Allah et al., 2019). This database subject of online open access, to meet the needs of actors working in the fields of linguistic, teaching, translation and communication.

Furthermore, it was also a question of endowing the Amazigh language with grammar tools. The first developed tool concerns the morphosyntactic tagger (Ataa Allah and Jaa, 2009), which helps linguists to annotate words, of the Amazigh corpus (Boulaknadel and Ataa Allah, 2013), as corresponding to a particular part of speech (Ataa Allah et al., 2013). The second tool concerns the Amazigh conjugator[8] (Ataa Allah and Boulaknadel, 2014b). It allows an online access to the conjugation of the Amazigh verbs. Other tools are under study, they concern morphological analysis (Ataa Allah, 2014; Nejme et al., 2016), named entities (Talha et al., 2015), spellchecker (Chaabi et al., 2019) and machine translation system (Miftah et al., 2017; Taghbalout et al., 2018).

4.4 Language Learning Materials

Language learning plays an important role in its preservation and the transmission of its culture. Aware of that, IRCAM has realized many digital teaching aids and educational materials[9], in order to contribute to the discovery of the cultural heritage and the development of the reading and the lexical competences. The objectives of these materials are to familiarize the learner with tifinagh, to support oral expression, to prepare and facilitate the transition to writing, to develop reading taste, and to enrich the learner's vocabulary.

The elaborated materials are structured on three kinds. Those for supporting reading: ⵜⵉⵣⵓⵣⴰⴼ [tizuzaf] 'Jewelry' e-book and the digital game ⴰⵏ ⵏⵎⵍⵏ ⵜⵉⴼⵉⵏⴰⵖ [Ad nlmd tifinaghe] 'Learning tifinagh'. Materials for developing vocabulary: the multimedia

[5]http://tal.ircam.ma/talam/ref.php
[6]http://tal.ircam.ma/talam/lexam.php
[7]http://tal.ircam.ma/dglai/
[8]http://tal.ircam.ma/conjugueur/
[9] tal.ircam.ma/talam

dictionaries ⵜⴰⵎⴰⵡⴰⵜ ⵏ ⵜⴰⵡⵍⴰⴼⵜ [Tamawalt inu tawlafant] 'My illustrated vocabulary' and ⵜⴰⵎⴰⵡⴰⵜ ⵏ ⵉⵎⵣⵢⴰⵏ [Tamawalt n imzzyann] 'Children dictionary' (Ataa Allah, 2011), the multimedia game ⵜⵉⵏⵎⵍ ⵏ ⵜⴰⵎⴰⵣⵉⵖⵜ [Tinml n tmazight] 'Amazigh School', and the mobile games ⵉⵣⵡⵍⵉⵏ ⴰ ⵜⴰⵎⴰⵣⵉⵖⵜ [Izwiln s tmazight] 'Numbers in Amazigh' and ⴰⵡⴰⵏ ⵏⵉ ⴰⵎⵣⵡⴰⵔⵓ [Awal inu amzwaru] 'My first words'. In addition to materials for developing communicative competence, including multimedia support: ⵉⵎⵓⵔⴰⵏ ⵉⴰⵎⵢⴰⵔⵏ ⴷ ⵉⵏⴰⵎⵢⴰⵔⵏ [Imudar iramyarn d inamyarn] 'Wild and Domestic Animals' and ⵉⴳⴷⴰⴷ ⴷ ⵉⴱⵓⴽⵅⴰ [IgDaD d ibukha] 'Birds and Insects'. In addition to these materials, IRCAM is actually working on a Massive Online Open Courses project that aims to make Amazigh language courses available for all interested persons inside and outside Morocco (Chaabi et al., 2018).

5. Conclusion

IRCAM has risen to a number of challenges, particularly those related to the Amazigh language integration in the digital world. Nevertheless, the validation of the organic law relating to the implementation of the Amazigh language as an official one, which stipulates its insertion into education and the priority areas of public life, opens up for IRCAM new challenges. In order to face these challenges, some research projects are underway and other are planned.

6. Bibliographical References

Andries, P. (2008). Unicode 5.0 en pratique, codage des caractères et internationalisation des logiciels et des documents. Dunod, France.

Ameur, M., Boumalk, A., Iazzi, E.M, Souifi, H. and Ansar K. (2009). Vocabulaire des médias, français-amazighe-anglais-arabe. IRCAM Publications, Rabat.

Ameur M., Bouhjar A., Boumalk A., El Azrak N. and Laabdelaoui R. (2009). Vocabulaire grammatical. IRCAM Publications, Rabat.

Ameur, M., Ansar K., Bouhjar A. and El Azrak, N. (2013). Terminologie amazighe de l'audiovisuel. IRCAM Publications, Rabat.

Ameur, M., Ansar K., Bouhjar A. and El Azrak, N. (2015). Terminologie administrative. IRCAM Publications, Rabat.

Ameur M., Ansar, K., Boumalk A., El Azrak N. and Laabdelaoui R. (2017). Dictionnaire général de la langue amazighe. IRCAM Publications, Rabat.

Ataa Allah, F. (2011). Conception d'un dictionnaire imagier sonore en ligne de la langue amazighe. In proceedings of the 8th Multidisciplinary Symposium on Design and Evaluation of Digital Content for Education (SPDECE'11), pages 158-164, Ciudad real, Spain.

Ataa Allah, F. (2014). Finite-state transducer for Amazigh verbal morphology. Literary & Linguistic Computing. Oxford University press.

Ataa Allah, F. and Boulaknadel, S. (2011). Convertisseur pour la langue amazighe : script arabe-latin-tifinaghe. In proceedings of the 2nd Symposium International sur le Traitement Automatique de la Culture Amazighe (SITACAM'11), Agadir, Morocco.

Ataa Allah, F. and Boulaknadel, S. (2014a). La promotion de l'amazighe à la lumière des technologies de l'information et de communication. Asinag, 9:33-48, IRCAM publications.

Ataa Allah, F. and Boulaknadel, S. (2014b). Amazigh verb conjugator. In proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14), pages1051-1055, Reykjavik. Iceland.

Ataa Allah, F., Boulaknadel, S. and Frain, J. (2019). Dictionnaire général de la langue amazighe informatisé. In proceedings of the conférence sur la Diversité Linguistique et TAL (DiLiTAL'19), Oujda, Morocco.

Ataa Allah, F., Boulaknadel, S. and Souifi, H. (2014). Jeu d'étiquettes morphosyntaxiques de la langue amazighe. Asinag, 9:171-184, IRCAM publications.

Ataa Allah, F. and Jaa, H. Etiquetage morphosyntaxique : outil d'assistance dédié à la langue amazighe. In proceedings of the 1st Symposium International sur le Traitement Automatique de la Culture Amazighe (SITACAM'09), pages 110-119, Agadir, Morocco.

Boukhris, F., Boumalk, A., Elmoujahid, E. and Souifi, H. (2008). La nouvelle grammaire de l'amazighe. IRCAM Publications, Rabat.

Boukous, A. (2012), Revitalizing the Amazigh language: stakes, challenges, and strategies. IRCAM Publications, Rabat.

Boulaknadel, S. and Ataa Allah, F. (2013). Building a standard Amazigh corpus. In Advances in intelligent systems and computing, 179:91-98.

Chaabi, Y., Boulaknadel, S. and Ataa Allah, F. MOOC-IRCAM pour l'apprentissage de la langue Amazighe - opportunités et défis. In proceedings of the 8th conférence internationale sur la Technologie de l'Information et de Communication pour l'Amazighe (TICAM'18), Rabat, Morocco.

Chaabi, Y., Outahajala, M. and Ataa Allah, F. (2019). Amazigh Spell Checker based on Naïve Bayes and N-Gram. In proceedings of the International Conference on Advanced Technologies and Humanitarian Sciences (ICATHS'19), Rabat, Morocco.

Commission Spéciale d'Education et de Formation (COSEF) (2000), La Charte Nationale d'Education et de Formation, Royaume du Maroc, Rabat, Morocco.

Miftah, N., Ataa Allah, F. and Taghbalout, I. (2017). Sentence-aligned parallel corpus Amazigh-English. In proceedings of the International Conference on Information and Communication Systems, Irbid, Jordan.

Nejme, F., Boulaknadel, S. and Aboutajdine, D. (2016). Amamorph: finite state morphological analyzer for Amazigh. Journal of computing and information technology, pages 91-110.

Talha, M., Boulaknadel, S. and Aboutajdine, D. (2015). Development of Amazigh named entity recognition system using hybrid approach. In proceedings of the 16th International Conference on Intelligent Text Processing and Computational Linguistics, pages 14-20, Caire, Egypt.

Taghbalout, I., Ataa Allah, F. and El Marraki, M. (2018). Towards UNL based machine translation for Moroccan Amazigh language. International Journal of Computational Science and Engineering, 17(1): 43-54.

Yakoubi, N., Frain, J. and Ataa Allah, F. (2016). Convertisseur Numérique : Tifinaghe-Braille. In proceedings of the 7th conférence internationale sur la Technologie de l'Information et de Communication (TICAM'16), Rabat, Morocco.

Proceedings, pages 219–222
Paris, UNESCO Headquarters, 5-6 December, 2019.

Language technology for Indigenous Languages: Achievements and Challenges

Sjur N. Moshagen, Trond Trosterud, Lene Antonsen
UiT The Arctic University of Norway
Tromsø, Norway
{sjur.moshagen, trond.trosterud, lene.antonsen}@uit.no

Abstract

Fifteen years of indigenous language technology development by UiT/Saami Parliament has resulted in spelling and grammar checkers, desktop/mobile keyboards, morphological analysers, MT, speech synthesis, language learning tools and intelligent electronic dictionaries. This was facilitated by an open source language independent infrastructure, targeted at languages with rich and complex grammar, with integration for host operating systems and apps. The current primary challenge is integration with closed platforms where we cannot currently support user needs. Our proposed solution is a "Manifesto for Open Language Technology", where APIs, localisations and source code are open, while ensuring community intellectual property custodianship, engagement and commitment.

Keywords: language technology, working solutions, morphology-rich languages

Čoahkkáigeassu (in North Saami)

UiT/Sámedikki 15 jagi eamiálbmot giellateknologiija barggu bohtosat leat sátne- ja grammatihkkadivvunprográmmat, boallobeavdi dihtorii ja mobiltelefovdnii, morfologalaš analysáhtorat, dihtorjorgaleapmi, hállansyntesa, giellaoahppanreaiddut ja intelligeanta digitála sátnegirjjit. Dát lea huksejuvvon rabas gáldokoda infrastruktuvrras, mii lea heivehuvvon gielaide main lea rikkes ja kompleaksa grammatihkka – infrastruktuvra mii siskkilda geavahanlavttaid ja applikašuvnnaid. Dál váldohástalus lea integreret prográmmaid giddejuvvon geavahanvuogádagaide, maid siste mii dál eat beasa doarjut geavaheddjiid dárbbsuid. Min evttohus lea "Rabas giellateknologiija manifesta", mas API:t, lokaliseren ja gáldokoda leat rabas, muhto seammás giellaservodagat galget hálddašit gáldokoda intellektuealla rivttiid.

1. Introduction

All indigenous languages of the world, except the Polynesian ones, are morphologically very complex languages. This means that one and the same word may show up in tens, hundreds or even thousands of forms. At the same time, most indigenous languages have a short written tradition and possess very small text collections, where the number of words in available running text cannot even be numbered in the thousands, let alone millions. Also, the text material there may often represent inconsistent or conflicting literary norms, and be of little use to language technology.

In this paper we present our the language technology infrastructure used to build LT tools for indigenous languages of the High North, languages with a rich and complex morphology. We also present a model for cooperation on the huge work behind language technology solutions that overcomes the problems posed by the weak commercial potential in such work. We will use the Saami language family as an example, but our model can be–and as a matter of fact has been–scaled to other indigenous languages as well.

The article is organised as follows: Section 2 gives some background of the Saami languages. Then we present our language technology and the methods used, an overview of remaining challenges, concerning problems of integrating indigenous language tools in mainstream computer platforms and programs. Finally comes a conclusion and a view on further work.

2. Background

In our work on language technology solutions we have focused on the Saami languages. Counting 9 separate languages, 8 of which have an official orthography, the Saami languages constitute the westernmost branch of the Uralic languages. The languages are spoken in the Mid and Northern part of the Scandinavian peninsula, the northern part of Finland, and the Kola peninsula. The largest of the languages is North Saami, with more than 25,000 speakers. All the other languages have less than 1000 speakers. South, Lule, Inari, Skolt and Kildin Saami have several hundred speakers, whereas Pite and Ume Saami have less than one hundred.

Typologically, the languages are unmistakably Uralic. They are suffixing languages with a rich nominal and verbal inflection, including person/number inflection on both verbs and nouns, different verb modes, as well as numerous derivational processes within and between the main parts of speech. Contrary to most Uralic languages they also have a rich variety of stem-internal morphophonological processes accompanying the suffixation, resulting in each paradigm possessing several inflectional stems. These processes includes the whole lexicon, and affects both root vowels and consonants, as well as stem consonants and suffix classes. The net result is that neither word form based approaches nor a system of stemming (suffix removal) is going to work.

Orthographically, each language has its own orthographical convention. Four of them (South, Ume, Pite

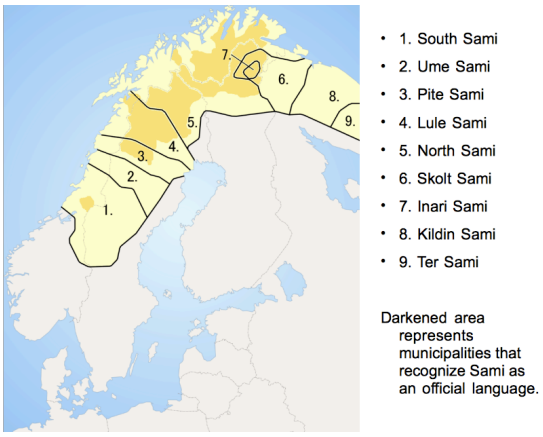


Figure 1: The Saami languages, and the municipalities where they have official status.

and Lule) build upon a tradition of writing the consonants according to the prevailing Scandinavian digraph tradition, and South Saami even has some vowel qualities similar to Scandinavian. The North, Inari and Skolt orthographies ultimately go back to a 200-year old tradition of one letter per phoneme, thus possessing a large repertoire of diacritical marks, each orthography still with its own conventions. Finally, Kildin Saami is written with the Cyrillic alphabet.

There are electronic text collections available for 5 Saami languages. North Saami is the largest, with 33M words, 3.5M parallel North Saami - Norwegian. The 4 languages all have less than 2M, from 200k Skolt Saami to 1.7M for Inari Saami. For the other Saami languages there are no online corpora available. (SIKOR, 2019). For an overview and a linguistic introduction to the language family, see (Sammallahti, 1998).

3. Our achievements

Fifteen years of indigenous language technology development by UiT/Saami Parliament has resulted in machine-readable grammars for most circumpolar literary languages, in the form of bidirectional models, capable of analysing and generating every word form of the language. These models in turn are used as key components in a wide array of tools, including spelling and grammar checkers[1], desktop and mobile keyboards[2], morphological and syntactic analysers[3], Machine translation[4], speech synthesis[5], language learning tools[6] and intelligent electronic dictionaries[7].

All the tools are in extensive use by the language communities. The spell checkers have been downloaded by approximately 2/3 of the language communities. On average, the e-dictionaries are used 12 times a week per speaker. The MT programs are in use in different contexts. Most notably, the Saami University College use our MT program to translate their web pages – being in North Saami only – into Norwegian.

The tools are discussed in several publications, a.o. on e-dictionaries (Johnson et al., 2013), spell checkers (Antonsen, 2018), grammarchecking (Wiechetek et al., 2019), Machine translation (Antonsen et al., 2017), and e-learning (Antonsen and Argese, 2018).

The grammar models we have made are made as bidirectional *finite state transducers*, as described in (Beesley and Karttunen, 2003). Grammatical ambiguities and syntax analysis we have resolved by means of Constraint Grammar (Karlsson, 1990).

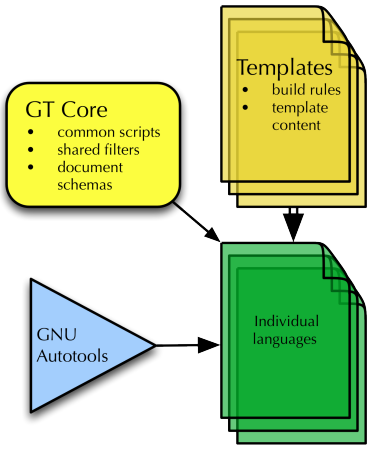


Figure 2: Our language independent infrastructure.

The challenge of scaling the work to several languages was handled by developing an open source, language independent infrastructure (see figure 2). It targets languages with rich and complex grammar, separating out the language specific work from language independent build and testing routines, and language agnostic integration into host operating systems and apps. In this way we have been able to reuse the work spent on integrating North and other Saami languages in office suites and other language processing software, thereby making such solutions available for language communities that simply do not possess the resources to achieve this by themselves.

A central part of our model is that it involves different stakeholders with different interests. For every language with an orthography, there is a community of university and field linguists devoting their career to it. For them, participating in making a machine readable model of the grammar is a way of empirically testing whether their hypotheses hold through. Philologists and lexicographers will find a way of formalising and testing their view on the vocabulary of the language. For language revivers and activists, the possibility of getting language technology tools is a central part of their strategy. The final part of this setup is the in-

[1] divvun.no/korrektur/korrektur.html
[2] divvun.no/keyboards.index.html
[3] giellalt.uit.no/lang/index.html
[4] gtweb.uit.no/mt/
[5] divvun.no/tale/tale.html
[6] https://oahpa.no/
[7] dicts.uit.no

frastructure for going from language model to practical program, and this infrastructure is provided by us as open source. The infrastructure is presented in (Moshagen et al., 2013).

4. Challenges

Currently, the primary challenge is integration with closed platforms where we cannot support user needs and meet their expectations. At present, language technology witnesses two conflicting trends. On one hand, more languages get some support or another by the major IT providers. On the other hand, the possibility of offering third-party solutions for language technology is restricted, often totally so.

In practice, **localisation of all major mobile operating systems is completely impossible**, and in practice also for all desktop operating systems and most apps. A language community has no way of defining or building their own digital presence, no tools to make their language visible and a natural part of the everyday language environment. This includes more things than the text on menus and buttons: it affects indexing and searching of text on computer systems, or hyphenation, needed for the long words resulting from complex morphology. For most of the world's languages it affects even such basic things as the name of the language: install a speller, and the name you get for that speller is not the actual language name, but a cryptic, three letter language code. How is an ordinary person meant to understand what that means?

Many operating systems, mobile and otherwise, provide a dictionary framework for adding dictionary content to the system. This is important for minority and indigenous languages. But often **those frameworks are not available to third parties**, or there is no way of adding lemmatisation or text analysis as part of the lookup process. For languages with complex morphology and phonology that is pretty much a blocker.

New web-based services and tools are a boon to many, but there is no support for languages outside the mainstream. A basic tool like a spell checker, which we have delivered for North Saami for 12 years now, is **suddenly locked out in new OS's like Chrome OS**, or web apps like Google Docs and Office 365. There is no way for us to provide it.

A great many indigenous and minority languages are using variants of the majority language alphabet, often with a lot of diacritics. The Unicode organisation has decided that they will not accept any new precomposed combinations of base characters plus diacritic, instead pointing to the mechanism for dynamic composition of diacritics. At the same time, this part of Unicode is not of great economical importance, since all majority languages are properly covered by precomposed letters in the standard. The end result is that **text written in minority and indigenous languages often becomes unreadable**, because the text rendering engines have bugs in them. The situation has been like this for more than 10 years (see figure 3).

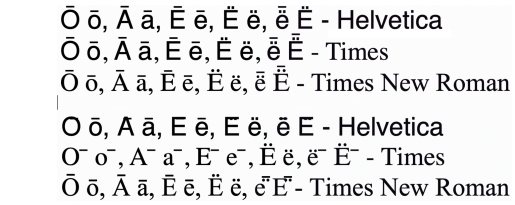


Figure 3: Kildin Saami letters as they should look, and as they often look. Notice how the accents in Helvetica has been dragged down into the base letter.

All the rage these days are machine learning and artificial intelligence, mostly applied on MT and speech technologies. These are exciting new opportunities, and the enthusiasm is impossible to miss. But the enthusiasm is hard to share from a minority perspective, for two reasons. The first and obvious reason is the resource demands required. Although the newest technologies do require smaller corpora and less text than the previous generation of machine learning, the demands are still way out of reach for most languages. But even if it would be possible, there would be very little point in doing it, due to the second reason: **all major operating systems save macOS are closed for speech services**. Dialog systems and virtual assistants even more closed, and the languages are not even known to the operating systems. All in all: **speech technology tools can never be used in practice.** They are nice demonstrations, and perhaps add to the body of studied languages, but in terms of tools for the user community the possibilities are slim.

Machine translation systems are more approachable, and there are programs for North Saami and a number of other languages. But often users wonder why Google Translate can't translate North Saami, and from a user perspective that is a legitimate question.

5. Our solution

How can we eliminate these technical hurdles? The issues being faced are not inherent to any language or writing system, but are **a consequence of technological considerations and economic incentives**. We propose a **"Manifesto for Open Language Technology"**, focusing on the following four points:

- **Open localisation:** all software should be localisable independent of the producer of the software
- **Open interfaces:** all language-related programming interfaces should be open by default
- **Open resources:** all language resources should be open and accessible for everyone, given the permission of the language community
- **Accessible standards:** all language-related international standards (ISO, etc) should be respected, fully implemented and implementations should be regularly updated

Immediate steps that can be taken by major vendors to **help us achieve indigenous self-determination in the digital realm** are as follows:

- **ISO 639 compliance** at the same rate as Unicode emoji compliance is supported in each major operating system
- Localisation packages for major operating systems should be installable from app stores and allow for **community-managed localisation**
- **Open up all language APIs currently held closed** on Windows, macOS, Android and iOS so that the community may integrate complex morphological tools with high quality user experience

The ultimate purpose it *not* that the major vendors implement language tools for us, but rather that there is a **guarantee of equal access to the APIs** that enable majority languages and allow them to be used for minority and indigenous languages.

We try to practice what we preach, by having all source code for the support infrastructure and tool integration on Github[8], building **tools in Rust to handle CLDR localisation data**, open source integration tooling for generating and maintaining **keyboards and locales for all major operating systems**, and developing continuous integration and continuous delivery infrastructure on top of the Azure platform.

The keyboard layout definitions and keyboard apps are also present on Github[9], with plans to migrate all remaining language technology source code from the currently used Subversion repository[10].

6. Conclusion

Indigenous languages need language technology made on their own terms. This may be implemented as a cooperation between university linguists and computational linguists, philologists and language activists, as well as programmers turning the language models into practical programs. In order for this to work the **software providers must open their software for third party providers**.

For minority and indigenous language communities, they need to have ownership over their own language, put their resources where they think it is most important, and **not be hindered by technical and economic decisions not related to their language at all**. It should be their decision whether they want their mobile phones and other devices to speak their language, not the decision of the vendor.

The ultimate goal is to achieve indigenous self-determination in the digital realm.

7. Acknowledgements

Thanks to our colleagues at Giellatekno and Divvun, as well as The Techno Creatives.

[8]https://github.com/divvun

[9]https://github.com/giellalt

[10]https://gtsvn.uit.no

8. Bibliographical References

Antonsen, L. and Argese, C. (2018). Using authentic texts for grammar exercises for a minority language. In *Proceedings of the 7th Workshop on NLP for Computer Assisted Language Learning (NLP4CALL 2018)*, pages 1–9. Linköping University Electronic Press.

Antonsen, L., Gerstenberger, C., Kappfjell, M., Rahka, S. N., Olthuis, M.-L., Trosterud, T., and Tyers, F. M. (2017). Machine translation with North Saami as a pivot language. In *Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa, 22–24 May 2017, Gothenburg, Sweden*, volume 29 of *NEALT Proceedings Series*, pages 123–131. Linköping University Electronic Press.

Antonsen, L. (2018). *Sámegielaid modelleren – huksen ja heiveheapmi duohta giellamáilbmái. [Modeling Saami languages. Construction and adaptation to real-world linguistic issues].* Ph.D. thesis, UiT The Arctic University of Norway, Tromsø.

Beesley, K. R. and Karttunen, L. (2003). *Finite State Morphology.* Studies in Computational Linguistics. CSLI Publications, Stanford, California.

Johnson, R., Antonsen, L., and Trosterud, T. (2013). Using finite state transducers for making efficient reading comprehension dictionaries. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013); May 22–24; 2013; Oslo University; Norway*, volume 16 of *NEALT Proceedings Series*, pages 59–71. Linköping University Electronic Press.

Karlsson, F. (1990). Constraint grammar as a framework for parsing running text. In *COLING '90 Proceedings of the 13th conference on Computational linguistics*, volume 3, pages 168–173, Helsinki.

Moshagen, S. N., Pirinen, T., and Trosterud, T. (2013). Building an open-source development infrastructure for language technology projects. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013); May 22–24; 2013; Oslo University; Norway*, number 16 in NEALT Proceedings Series, pages 343–352. Linköping University Electronic Press.

Sammallahti, P. (1998). Saamic. In Daniel Abondolo, editor, *The Uralic Languages*, pages 43—96. Routledge, London.

SIKOR. (2019). UiT The Arctic University of Norway and the Norwegian Saami Parliament's Saami text collection, with grammatical analysis. `http://gtweb.uit.no/korp/`.

Wiechetek, L., Moshagen, S., Gaup, B., and Omma, T. (2019). Many shades of grammar checking – launching a constraint grammar tool for north sámi. In Eckhard Bick et al., editors, *Proceedings of the NoDaLiDa 2019 Workshop on Constraint Grammar: Methods, Tools and Applications, Turku, Finland*, volume 33 of *NEALT Proceedings Series*, Linköping, Sweden. Linköping University Electronic Press.

Proceedings, pages 223–226
Paris, UNESCO Headquarters, 5-6 December, 2019.

Language Technology Applications in Africa within an "Inclusive, Innovative and Reflective" Crisis

Evelyn Fogwe Chibaka
Department of Linguistics, Faculty of Arts
University of Buea, Cameroon

Abstract

Any valid and objective contemporary scientific finding must be appropriately researched and made easily accessible via digitalized technology and techniques for public consumption. "Inclusiveness, innovation and reflectiveness" pose a crisis interface environment in the implementation of "Language Technology" in Africa; thus it will entail us into examining each of the variables and their corresponding effects. So far, "There is NO African solution to African problem fix here!" Let us put into practice the UN's "Leaving no one behind" motto, which is an integral part of the 2030 Agenda for Sustainable Development. Keywords: language technology, inclusion, innovation, reflection, Crisis interface.

Keywords: language technology, inclusion, innovation, reflection, crisis interface

Semkwep nkwel

Kikii dinyi i len nunu ditéé, yokiyo yi iyôni, ni i ihéli, inlama nyégsaba le iwéna, ikôda, itjéma ni manjel ma bôt banla pala jubul, tôbôtôbô ma lôak i ngambi i minsongi, ni imakenge mankiha ni mo. Ingéda dimbénge kikii likenge li ngambi minsongi inyu mahop inyep ni njômbi ikot bôt, to ni njômbi i mahol, ndi to njômbi ihek béngba djomede, dinléba le, lini likenge linla bé pala hol i kéndi yés i Afrika; hala antinde bés iwan manjom mantuga mana mahol. Ibol ingéda ini ditéé, Afrika añéba bé su le ayé le ala léba dipa inyu litôngi jéé munu ini njel; inyu kii? Inyule to Likoda li Maloñ ma Africa (UA), to ACALAN nu ayé hikuu hyé hikeñi hi mbénge mahol ma mahop, bankômôk bé dipa inyu ndôngi munu ini njel. We, sañ ijubga ni ni nguy inyu iyônôs ndak i Likoda li Makoñ ma nkoñ hisi (ONU) impot le, "tomut ayéglege bañ i mbus", ndak iyé le iyé yada ikédé makak makeñi ma bikéga inyu mapuhul ma maloñ momasô inwaa nwii 2030 inkola. Miño mi bibuk: likañ li yigil i mahop, likot, ligwal, kék i ndudu, bitéé.

1. Introduction

Considering the *indispensible nature of language* in almost all spheres of human existence or interaction on earth, the description, documentation, sustainability, revitalization, and usage of language has been made practical in facilitating our day-today activities via the current evolving insights of *Language technologies* and their applications. The question is –"Is the above statement of FACT a truism to African academia?"

In Africa, less than 20% of the rich Indigenous Knowledge (IK), researched data and/or findings have been digitally catalogued and barely 1-2% is imaged. This renders information highly inaccessible and underused especially because of the remoteness from or the poor networking with and/or applications of modern technologies and advance research techniques infrastructures.

2. The Crisis Interphase

A **crisis** in this context is considered as "a turning point or situation in which there are a lot of problems that must be dealt with *quickly* so that the situation does not get worse or more dangerous → **emergency!"** An **interface** here refers to "the common boundary between two things – the point of interconnection between entities; or it could be 'the place where things happen". (**https://d4htechnologies.com**)

2.1 Inclusiveness

Africa as a continent is in a crossroad as a passive or an observational participant in the global engagement of "Language technology application". Poor Governance and prioritization mix-up have deprived the continent from focusing on the advantages of digitizing and providing access to huge data resources that require technological, socio-cultural, and organizational capacity enhancements across the continent.

Most of our African languages are not yet supported with the modern technological expertise trend (e.g. Smart keyboards, Speech technology needs, Apps, etc.). So how inclusive

can Africa be in the global move into the language and technologically driven world, when there is no enabling environment for the African scholars to effectively contribute or be part of the game?

2.2 Innovation

The Western World is fast developing constructive clusters of collaborative "Research Infrastructures, integrating initiatives, e-infrastructures and other world-class infrastructures, and building bridges between different although tightly interrelated fields, etc." This is in a bid to adequately address Social Sciences and Humanities data lifecycle's management; even though most African scholars/researchers apart from South Africans with their SADiLaR[1] and a few pockets in the Western (ALORA[2]), and Eastern countries are still to be aware of or affected with the current waves of "Language technology" innovation fever.

For example PARTHENOS[3] is making great strides in strengthening the cohesion of researchers and their rich massive data archives from the Social Sciences and Humanities to generate a broad-spectrum e-infrastructural pool that will build bridges and further provide a resource bank for resolution of challenging societal demands. Dynamic bodies like CLARIN[4], DARIAH[5], etc. are seriously reforming and transforming research technology capacity building platforms, data management policies, awareness on research data repositories, etc.; all geared towards bringing innovative strategies into New Knowledge through language technology management. This however makes it possible for oral and written language, cultural heritages and digitized data to be conveniently converted into problem solving ventures.

In most of African countries, there is a crisis of digital technologies and social media applications to create and share our language data resources or even convert them to address our societal issues. Most Africans are reduced to receivers or consumers of language technological soft wares and Applications. Regrettably, most of them are not yet involved in the designing or creation of these research techniques or actively engaged in the race for rapid expansion of access to digital technologies in the global digital landscape. At this point, how do the African humanities scholars significantly contribute in bringing innovation in the sphere of language technology without dynamic collaborative organizations or projects like those in the West to empower and motivate them to recreate and/or bring out their God endowed talents?

2.3 Reflection

Africa's complex linguistic and cultural specificities do not make the language technology and data management (policies) any easier for her researchers. Africa is confronted with a generation of highly unacquainted and unskilled technology-driven researchers in the same platform with more experienced and fast lane technologically apt researchers in the Western World without complicated multilingual and multicultural diversities.

In reflecting about our African cultural heritages, we are placed in a position of struggling to generate new ideas, strategies and governance structures. These constitute the basis for overcoming the numerous crisis in Africa (i.e. economic innovations, socio-political stability, multilingual/cultural heritages diversity integration with emerging technologies and techniques, absence of basis subsistence facilities, poor infrastructures, geographically enclave suburbs, health challenges, etc.) to foster development and a better living. Thus, with all these internal and external vices, it becomes a major problem breaking through our existing (most often complicated) multicultural and multilingual diverse ideologies. Even though this is necessary in order to create a Pan-African

[1] **SADiLaR** "South African Centre for Digital Language Resources" is arguably the leading hub in 'Language technology' advancement in Africa. It focuses on all official languages of South Africa, supports research and development in the domains of language technologies and language-related studies in the humanities and social sciences. The Centre facilitates the creation, management and distribution of digital language resources, as well as applicable software. SADiLaR has developed machine translation engines for all the eleven official languages of South Africa.

[2] **ALORA** "Archives of Languages and Oral Resources of Africa" is a project created and run by the "Centre international de recherché et de documentation sur les traditions et les langues africaines" (CERDOTOLA) member states. It is a virtual infrastructure for hosting versatile digital resources stemming from documentation of African languages and cultural heritages. It is hosted and managed in Yaounde, Cameroon.

[3] **PARTHENOS** stands for "Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies".

[4] **CLARIN ERIC** is the governing and coordinating body of CLARIN (Common Languages Resources and Technology Infrastructure). It is a consortium of ca. 20 countries and intergovernmental organizations (and counting), represented by their ministries Content. Their sole objective is in providing Social Sciences & Humanities access to digital language data & advance technological tools.

[5] **DARIAH** is a pan-european infrastructure for arts and humanities scholars working with ICT-enabled methods. It supports digital research as well as the teaching of digital research methods.

language technology network or organization that will have functions, projects, and infrastructures like CLARIN, DAHRIN, or PARTHENOS, etc.

Africans are at a turning point. That is, how to move from the current stalemate in Africa (i.e. with negatively polarized challenging environments) to having African humanities scholars connecting, contributing their distinctive worldviews and perspectives in debates and contemporary research. Such debates of course should be towards evolving "language technology" to impact on sustainable language and societal or global development with the existing and almost virgin indigenous knowledge resources.

3. Recommendations/Strategies

The language technology crisis interface in Africa has to be addressed at two levels: the Individual and the Institutional. This is in order **not** to be left behind in this technological advancement vision or race.

3.1 Individual Strategies

We the African scholars first have to take off from a change of "Mind-set". We need a mind-set that nurtures the necessity of actively engaging in language technology, trainings, projects, and partnership collaboration with other scholars (both national and international).

Individual efforts must be made by the African scholars to acquire basic computational competence, technological research methods, techniques and Applications, seek for capacity building workshops, conferences, Summer schools, online networking programs, etc. that will empower, introduce, and expose them to the digital generation.

African scholars should take up the initiative into the creation of a Pan-African Language technology infrastructures Association (just like what some other African Humanities scholars introduced in Leiden this July 2019). We can only achieve assistance from international bodies if we take the first step. There are already many landmark existing Organizations in the Western World that serve as templates and even mentors (and/or sponsors) for such innovative initiatives.

At individual levels, the African scholars should be involved in critical thinking, creativity and culturally sensitive, methodological, and technological software/Apps for the development of African IK resources. North-South technological and techniques exchanges and transfer can achieve this. Thus, constant and close intellectual collaboration and networking will make significant progress in our language technology endeavours.

Today we are in a global village where the lack of the knowledge of information technology to an extent is a grievous crime. In this respect, African scholars should put in more efforts into surfing and canvasing for competitive international grants or projects that are language technologies oriented; through this they will become team players in the technology game.

3.2 Institutional Strategies

The good governance practice in our African nations will play a very vital role as "Game changers". This will require a major change of 'mind-set' and cultivation of the good WILL on the part of the government to develop technological infrastructures and make substantial provisions for favourable virtual environments in institutions or Universities. These institutions will subsequently develop data management systems/structures for more international (N-S) 'language technology' cooperation and collaborative networking that should be sponsored and meticulously monitored.

Our governments should provide subsidies for 'Language technology' workshops, trainings, conferences, Summer schools organizations, etc. The provision of pooling activities, resources and tools for heritage e-research networking, optimization and synergies will make all the difference in our continent.

The inclusion of language technology programs in the University syllabus will play a great role in show-casing the programme and attracting many more scholars to specialize in this academic exercise. In addition, the universities can significantly go a long way to expanding research infrastructures visibility such that strategic inter-university partnerships, cooperation links or research collaborations are promoted and strengthen.

The development of the society or our different African nations will greatly re-enforce the prevalence of enabling research environment since the availability and stable **Electricity** and **Internet** supplies are basic requirements for effective technological operation.

If the African states can facilitate a) the development of new collaborations, b) user accessibility, c) collaborative and innovative actions, d) funding opportunities, and, e) knowledge transfer opportunities in their different institutions or infrastructures, the continent will leap out of its crisis interface.

4. Conclusion

There is language technological work force everywhere in the world. If given the opportunity and/or enabling environment they will develop to meet up with our varied human needs. This is

especially because we are all humans even though our challenges or differences are superficial and defined by our unique geographical and socio-cultural specificity, which we have to consider when designing, employing, interpreting and disseminating roles of language technologies and software Applications. If technological developments are not culturally sensitive (or adapted), its unconscious applications will definitely not generate any guaranteed or expected research Outcomes. We therefore operate with the formula that **TECHNOLOGY + CULTURE ⇨ EXPONENTIAL DEVELOPMENT.**

In a nutshell, the inclusion, innovation and reflection on the language technology applications in Africa with its peculiar geographical, socio-political, economical, multicultural heritage, and governance complexities can be arrested. This can be done through the implementation of the 'Crisis Now' management model for Africa; that is, an urgent need for more collaborative interactions between language technology experts with existing major bodies like the AU and ACALAN[6] to operate on innovative engagement exchanges, and information management strategies that will greatly project language technology infrastructures development in Africa.

Using this medium, I wish to appeal to international Organizations to help African Scholars in their partnership collaborative networking to get them into the emerging Language technology (e-infrastructure) world i.e. putting Africa into their technological evolutional dream Agenda, projects and research programs. Together as a **willing team**, the Western and the Southern world will make this our environment a paradise on earth.

5. Acknowledgements

I heartily thank the organizers of LT4All 2019 conference for inviting and sponsoring my participation in this one of its kind international think-tank event.

6. Bibliograhical References

European Commission. (2014). Horizon 2020, The EU Framework Programme for Research and Innovation, Societal Challenges, http://ec.europa.eu/programmes/horizon2020/en/h2020-section/societal-challenges, (accessed on 15 November 2019)

Intro to CLARIN, https://www.clarin.eu (accessed on 14 November 2019)

Working with parliamentary records: https://www.clarin.eu/event/2017/clarin-plus-workshopworking- (accessed on 14 November 2019)

SRIA Editorial Team. (2017). Language Technologies for Multilingual Europe – Towards a Human Language Project (SRIA Version 1.0) http://www.cracking-the-language-barrier.eu (accessed on 15 November 2019)

ParlaCLARIN @ LREC: https://www.clarin.eu/ParlaCLARIN (accessed on 14 November 2019)

ParlaFormat: https://www.clarin.eu/event/2019/parlaformatworkshop (accessed on 14 November 2019)

[6] **ACALAN** is 'The African Academy of Languages', which is an arm of the 'African Union' (AU) whose major role is to foster Africa's integration and development through the development and promotion of the use of African languages in all domains of life in Africa.

Proceedings, pages 227–230
Paris, UNESCO Headquarters, 5-6 December, 2019.

Digital Surveillance and Digitally-disadvantaged Language Communities

Isabelle A. Zaugg, PhD
Postdoctoral Research Scientist, Data Science Institute, Columbia University
Northwest Corner, 550 W 120th St #1401, New York, NY 10027, United States
iz2153@columbia.edu

Abstract

The issue of digital surveillance often falls outside urgent discussions regarding the need to build digital supports for under-resourced languages. While the benefits of these supports for digitally-disadvantaged language communities are clear, the reality is that standardized script use, standardized spelling, and NLP systems in particular increase a language community's legibility for digital surveillance. As we build digital supports for Indigenous and minority language communities, we must consider how these tools might be used against them through digital surveillance, and how to combat these risks.

Keywords: digital surveillance, language diversity, Indigenous communities

ፅንሰ ሀሳብ

ይህ ጥናታዊ ፅሁፍ በድጅታል ድጋፍና በቋዋንቋዋ ብዝሐነት ዙሪያ በሚደረጉ ውይይቶች ውስጥ አዝወትሮ ስለ ማይነሳው ስለዲጅታል ክትትልና ስለላ ጉዳይ ያትታል። የቋንቋዎች ዲጅታይዜሽን ጥቅሞች መጠነሰፊ ቢሆኑም በሌላ በኩል ደግሞ የነዚህን ቋንቋ ተናጋሪዎች ላልተጠበቀ አደጋ ሊያጋልጣቸው ይችላል:: በዚህ ፅሁፍ እኒህ አደጋዎች እንዴት ሊከሰቱ እንደሚችሉና እንዴት ልንከላከላቸው እንደምንችል ሀሳቦች ይቀርባሉ።

1. Introduction

This paper shines a light on an issue that lingers just outside many discussions on topics of digital supports for language diversity, digital surveillance. This is an issue that has increasingly caused me concern as a scholar and advocate for improved supports for digitally-disadvantaged languages, in particular the Ethiopian and Eritrean languages written in the Ethiopic script. In this text I raise issues and questions arising out of the field of critical data studies to pose what I hope will be a productive challenge to the participants of this conference as we focus on serving the digital needs of speakers of diverse global languages. While our work is done under the banner of equity and appreciation for the rich history and culture contained within each language tradition and its community of speakers, we should not fall prey to a blind techno-optimism that contends that we can find purely technical solutions to entrenched and troubling problems of social inequality and injustice. We must consider the way in which the very tools that can bring important benefits to language communities can also be turned against them/us. This paper explores a number of issues, and poses a series of questions, at the intersection of digital surveillance and digital tools focused on the under-resourced languages of minority and Indigenous communities.

2. Context

Many minority and Indigenous language communities, as well as the speakers of regionally-dominant and national languages that have not been target markets for global tech companies, are eager to see the development of digital tools that support their languages. The reasons are crystal clear. Without these tools, when using digital tech to read, write, and edit text, users' options are highly limited. They can choose to communicate in a globally dominant language that is well-supported, or transliterate their messages into a dominant script such as the A,B,C's of Latin. Alternatively, they can use workarounds like sending images of hand-written text or short audio "text messages." Or they may forego the use of digital tools, losing out on potential benefits, a language-induced digital divide. Or perhaps they may become motivated to develop their own digital supports such as a proposed Unicode encoding of their script, an easily accessible and well-designed keyboard, a font, spellchecker, a translated social media interface, etc… (Zaugg, 2019).

Many advocates and designers of such digital tools for under-resourced languages are motivated by the hopes of keeping their language and language community vibrant in the face of linguists' predictions that 50-90% of languages face extinction this century (Harrison, 2007; Kraus, 1992). If a language can achieve a digital foothold, the hope is that young "digital natives" will not forego their mother tongue under the impression that other more dominant languages are cooler, more modern, and more convenient for the digital sphere and wider life (Rehm, 2014).

The stakes are high and the benefits of digital inclusion clear. And considering that under-resourced languages have never been and are unlikely to become a major focus of large tech corporations, there is a lot of work for volunteers and passion-driven advocates and technologists to do. But what are the risks and drawbacks, if any, of bringing a language, particularly minority and Indigenous languages, into the digital sphere? This paper focuses on the issue of digital surveillance, and the role it may play in complicating this picture of digital solutionism for language communities that often face unique and complex vulnerabilities, as well as unique resilience.

Digital surveillance can take a number of forms, from the sometimes forced collection of biometric data (Wee, 2019), CCTV systems that monitor the behavior of a city or country's residents (Diamond, 2018), and the monitoring of verbal and written communication, which typically rely on the legibility of a language to computers. This means that the more advanced digital supports and NLP systems

are for a language, the more transparent the community using that language becomes to powerful forces that wish to surveil them. This is of particular import for us to consider, as the minority and Indigenous languages that are least supported digitally are also disproportionately at odds with national governments and corporate powers, sometimes by virtue of their very existence. These communities may also be particularly vulnerable to target marketing promoting the most injurious aspects of global modernity, and are often spread globally in diasporic patterns that may increase their digital communication needs.

In one telling example, we see Chinese-speaking #metoo activists sending text messages containing hand-written characters photographed upside down in order to pass the censorship of OCR-based (optical character recognition) AI systems designed to stifle their dissent (Weerasekara, 2018) – the very opposite of the types of standardized and legible digital tools we are promoting here. Therefore, it is critical that we consider the extent to which digital tools represent a double-edged sword for the communities we hope to serve, and think actively about the ways in which to harvest digital benefits while guarding against their vulnerabilities.

3. Military-intelligence Surveillance

While the vast majority of the approximately 7000 languages of the world are digitally under-resourced, in some case tools to work with these languages exist in the private domains of military-intelligence projects. These tools are built in order to surveil and in some cases constrain the activities of groups that pose national security concerns, who in many cases are also members of minority and Indigenous language communities. These tools are often developed by resource-rich countries with a high focus on military-intelligence (European Commission, 2006; "IARPA MATERIAL Program," 2017). It stands to reason that in some cases these tools are developed by or have the potential to fall into the hands of authoritarian governments or corporate entities whose interests are at odds with the language communities in question.

It is important to recognize that some minority language communities contain groups that propose violent means to achieve separatist or supremacist aims, and surveillance of their activities is essential to saving lives. But in turn, other language communities pose a threat to oppressive regimes by their simple existence, such as entrenched plutocracies that wish to clear-cut rainforests populated by Indigenous peoples who have been the historical residents of these areas and are also guardians of their biodiversity (Muñoz Acebes, 2019). Unfortunately, it would be naïve to think that language tech "for all" is a simple good, untainted by the same power structures that have left these languages digitally under-resourced in the first place. I hope that a concern for the human consequences, both intended or unintended, of the technologies being promoted at this conference under the banner of equity will be a highlight of conversations throughout the conference.

A question worth posing is: To what degree are NLP innovations resulting from high-investment research on under-resourced languages carried out in the military-intelligence sector eventually made available to serve the needs of language communities (if any)? This would not be unprecedented – for example, consider DARPA building the backbone of the Internet, then making it available for public use (Cerf, n.d.). An additional question: To what degree does military-intelligence research feed off of open-source corpora and data sets provided by academics and language advocates in the hopes of pooling resources to build better tools for their communities? What is the direction of funding, data, and tools developed in this arena, and whose needs do they serve?

4. Surveillance Capitalism

Digital surveillance is no longer the exclusive purview of traditional agents of surveillance, such as governments. On the contrary, digital technologies make surveillance, or "big data analysis" as it is often euphemistically termed, an activity available to almost any actor that can pay. In her 2019 book, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*, Shoshana Zuboff builds upon the work of many critical data scholars to propose the term "surveillance capitalism" as the defining economic structure of our era. Surveillance capitalism, she proposes, uses human experience, mediated by digital and "smart" technologies and often extracted without consent, as a free raw material that can be translated into behavioral data. This behavioral data is fed to machine learning systems that provide predictions about what people will do in the future. She documents how surveillance capitalists have gained immense wealth through the trading of "prediction products," as companies profit from laying accurate bets on people's future behaviors. These systems tend to reward the privileged while entrapping the underprivileged, whose choices are particularly constrained.

In the context of this conference, we should consider how the social impacts of surveillance capitalism, will impact minority and Indigenous language communities, many of which already exist in economically and socially precarious positions. We know that online systems that market themselves as free and useful tools are designed to be addictive, through such strategies as "infinite scroll," and to promote superficial values and capitalistic consumption of advertised goods (Center for Human Technology, n.d.). And we know that platforms like Youtube push sensational content to keep users on their platforms, making them the perfect breeding ground for actors seeing to radicalize youth and divide communities (Fisher & Taub, 2019). Therefore, we need to think not only about building robust supports so under-resourced language communities can "join" these systems, but also how they can protect themselves once they are there. Or more radically, how can we dismantle the negative sides of these systems while preserving their benefits, to enhance the lives of all, so that as members of under-resourced language communities may choose to join these global domains, they find themselves in digital spaces that honor and enhance their lives?

5. Humanitarian Surveillance

A push for language technologies often takes on a sense of urgency during humanitarian crises. The need is clear - digital technologies can help connect displaced communities or spread life-saving information during

ongoing disasters. Yet, we must also think about how to address long-term risks already vulnerable language communities may face in the context of tech solutions for humanitarian crises.

Mirca Madianou (2019) has demonstrated that digital innovation and data collection practices are increasingly core components of humanitarian response, yet in the long-term tend to further entrench discrimination and power asymmetries that disadvantage affected populations. This takes place, for example, when data first collected to identify, serve, and give voice to refugees, later, through "function creep" (Ajana, 2013), is used to monitor refugees' activities and limit their movements as their status shifts from objects of pity to national security threats (Madianou, 2019). Sensitive data is often collected through partnerships between humanitarian organizations, large corporations such as Accenture, Google, Microsoft, etc., and governments with whom the UN works hand and glove (Madianou, 2019).

These multi-faceted partnerships raise a number of questions about who owns, profits from, and can access this data in the long-term. For example, Rohingya refugees have expressed grave concerns that personal data collected by humanitarian organizations may be shared with the government of Myanmar, the same actor that perpetrated atrocities against them (Madianou, 2019). On the corporate side, too often the interests of vulnerable populations are forgotten when their data can be put to other uses, such as to improve facial recognition systems (Madianou, 2019) that may be sold back to governments seeking to keep refugees out. How might improving digital supports for the languages of humanitarian aid recipients potentially open their communities to further harms?

6. Positive Surveillance : Content Moderation

Paradoxically, when considering potential harms of digital surveillance, we must also consider the need for and persistence gaps in positive forms of surveillance and oversight. Content moderation, ideally community-led, is essential on open digital platforms in order to preserve the safety of these spaces, particularly for vulnerable communities. For example, Facebook serves around 1/3 of the world's population, with more and more language communities represented among its membership, yet its content moderation for all but the most globally dominant languages is close to non-existent (Koebler & Cox, 2018). As incidents of violence stemming from hate speech and fake news in Myanmar (Samet, 2018), Nigeria (Adegoke & BBC Africa Eye, 2018), and elsewhere demonstrate, this gap has life and death consequences.

Social media platforms like Facebook are ill-equipped to combat these trends as they simply have not invested the resources in personnel and AI systems that understand local languages and social tensions at play. Their business models raise questions as to whether this trend will be reversed without a major paradigm shift. Consider that the data generated by users in developed contexts like the U.S. is far more monetizable today than data generated by the company's huge and growing user base in the linguistically-diverse developing world. Furthermore, in a legal context in which even the U.S. government is struggling to hold Facebook, Google, and Twitter to account for their role in foul play perpetrated on their platforms (Bergen, Frier, & Wang, 2017), it seems unlikely that a developing country could succeed in this regard today.

Nigeria provides one stark example of how Facebook's gaps in language awareness and content mediation can lead to deadly violence and escalating tensions. After the June 2018 circulation of photos of graphic violence – viewed tens of thousands of times on Facebook - combined with false statements about an ongoing massacre in Plateau State, 11 people were killed in retribution in a town several hours north (Adegoke & BBC Africa Eye, 2018). In response, Facebook promised to strengthen its moderation of Nigerian content, a country with 24 million monthly Facebook users in 2018, and where 53 million Internet users are predicted to come online by 2025. But when BBC Africa Eye investigated Facebook's new "third-party fact-checking program," they found just four full-time employees in Nigeria to analyze and take-down fake news, and none of them speak Hausa, a language spoken by millions in Nigeria (Adegoke & BBC Africa Eye, 2018). For those of us promoting digital supports for digitally-disadvantaged languages, how might we also advocate for the human supports needed to make digital spaces safe for their speakers?

7. Conclusion

How can we work with digitally-disadvantaged communities to balance both the goods and harms of digital supports? This paper, I acknowledge, offers more questions than answers. Certainly, infrastructure must be considered, as backdoors to data may be built into systems at the outset of digital infrastructure developments (Aglionby, Yang, & Feng, 2018). End-to-end encryption, trusted intermediaries, and community oversight and moderation are also essential. Community networks based on a commons model might also offer bespoke solutions for some language communities. I hope experts in data security and data sovereignty will weigh in on these questions, in conversation with language communities themselves and their digital advocates, spurred to thought by the questions raised here.

8. Bibliographical References

Adegoke, Y., & BBC Africa Eye. (2018, November 13). Like. Share. Kill. Nigerian police say "fake news" on Facebook is killing people. *BBC News*. Retrieved from https://www.bbc.co.uk/news/resources/idt-sh/nigeria_fake_news

Aglionby, J., Yang, Y., & Feng, E. (2018, January 29). African Union accuses China of hacking headquarters. *Financial Times*. Retrieved from https://www.ft.com/content/c26a9214-04f2-11e8-9650-9c0ad2d7c5b5

Ajana, B. (2013). *Ajana: Governing through biometrics: The Biopolitics of Identity*. Retrieved from https://scholar.google.com/scholar_lookup?hl=en&publication_year=2013&author=B.+Ajana&title=Governing+through+biometrics

Bergen, M., Frier, S., & Wang, S. (2017, October 10). Google, Facebook, Twitter Scramble to Hold Washington at Bay. *Bloomberg*.

Center for Human Technology. (n.d.). The Problem. Retrieved November 30, 2019, from Center for Humane Technology website.
Cerf, V. (n.d.). A Brief History of the Internet & Related Networks: Introduction. *Internet Society*.
Diamond, A. M., Larry. (2018, February 2). China's Surveillance State Should Scare Everyone. *The Atlantic*.
European Commission. (2006). *Human Language Technologies for Europe*.
Fisher, M., & Taub, A. (2019, August 11). We Wanted to Know How Online Radicalization Was Changing the World. We Started With Brazil. - The New York Times. *The New York Times*.
Harrison, K. D. (2007). *When Languages Die: The Extinction of the World's Languages and the Erosion of Human Knowledge*.
IARPA MATERIAL Program. (2017, August 17). *National Institute of Standards and Technology, U.S. Department of Commerce*.
Koebler, J., & Cox, J. (2018, August 23). The Impossible Job: Inside Facebook's Struggle to Moderate Two Billion People. *Vice*.
Kraus, M. (1992). The World's Languages in Crisis. *Linguistic Society of America*, *68*(1), 4–10.
Loomis, S. R., Pandey, A., & Zaugg, I. (2017, June 6). Full Stack Language Enablement. Steven R. Loomis website: http://srl295.github.io/
Madianou, M. (2019). Technocolonialism: Digital Innovation and Data Practices in the Humanitarian Response to Refugee Crises. *Social Media + Society*, *5*(3).
Muñoz Acebes, C. (2019, November 14). The Amazon Rainforest's Defenders Are Under Attack in Brazil. *Foreign Policy*.
Rehm, G. (2014). Digital Language Extinction as a Challenge for the Multilingual Web. *Multilingual Web Workshop 2014: New Horizons for the Multilingual Web*. Presented at the Madrid, Spain. Madrid, Spain: META-NET.
Samet, O. (2018, April 20). Assessing Facebook's Role in the Violence Against the Rohingya. *Pacific Standard*.
Wee, S.-L. (2019, February 21). China Uses DNA to Track Its People, With the Help of American Expertise. *New York Times*.
Weerasekara, P. (2018, April 27). "Historic moment": China's #MeToo activists use blockchain to skirt censors. *Hong Kong Free Press HKFP*.
Zaugg, I. (2017). *Digitizing Ethiopic: Coding for Linguistic Continuity in the Face of Digital Extinction* (Doctor of Philosophy in Communication, American University).
Zaugg, I. (2019). Imagining a Multilingual Cyberspace. In *Next Generation Internet*. *Finding ctrl: Visions for the future Internet*. https://findingctrl.nesta.org.uk/

Proceedings, pages 231–234
Paris, UNESCO Headquarters, 5-6 December, 2019.

Language Resources and Tools Development for Indonesian Languages

Budi Irmawati[1*], Arik Aranta[1], Wirarama Wedhaswara[1]
M. Iqbal D. Putra[1], Siti Oryza Khairunnisa[2]**
[1]Universitas Mataram, Jl. Majapahit 62 Mataram, Indonesia
[2]Tokyo Metropolitan University, 1 Chome-1 Minamiosawa, Hachioji, Tokyo 192-0397, Japan
{budi-i, arik, wirarama}@unram.ac.id, iqbaldwinfor@gmail.com, siti-oryza-khairunnisa@ed.tmu.ac.jp

Abstract

These works generated resources for languages in Indonesia. We started our works on Indonesian and Balinese and will continue with languages in the west part of Lesser Sunda Islands. We collected parallel learner sentences, documents with different age levels, and scientific papers. We use those resources to solve some problems such as preposition error correction, identify words for different age levels, and mapping reviewers who best match with a submitted paper in a referred publication. Then, to preserve ancient documents, we defined an input mechanism to write Balinese scripts recognized as the *Bali Simbar* font.

Keywords: Age level word list, Balinese transliteration, Dependency annotation scheme, Indonesia L2 resources

Résumé

Tulisan ini menjelaskan kegiatan-kegiatan untuk mengumpulkan data dan membuat aplikasi dalam bahasa-bahasa di Indonesia untuk membantu masyarakat dalam mempelajari bahasa-bahasa tersebut. Kami mulai dari bahasa Indonesia dan bahasa Bali dan akan melanjutkan dengan bahasa-bahasa di Nusa Tenggara Barat. Kami mengumpulkan kalimat paralel yang terdiri dari kalimat yang ditulis oleh pelajar bahasa, documen dari berbagai tingkatan usia, dan paper akademik. Kami menggunakan sumber daya tersebut untuk menyelesaikan beberapa permasalahan seperti perbaikan kata depan, mengenerate kata-kata yang digunakan pada tingkat usia tertentu, dan menentukan reviewer yang bidang keahliannya sesuai dengan paper yang disubmit ke sebuah seminar. Selanjutnya untuk menjaga keberlanjutan bahasa daerah, kami membuat metode untuk menuliskan huruf dalam bahasa daerah, dalam hal ini huruf Bali yang telah dikenal sebagai Simbar Bali dalam format UTF.

1. Introduction

Many people agreed that producing language resources are time consuming and highly cost. In the case of under-resource languages, large raw texts are also difficult to find. The problems are even harder because rare institutions that want to financially support those productions. Many people do not understand that language resources are really worth. Not many people know that recent methods may generate automatic decision extracted from a large sized of text data especially ones that have already been annotated, as those happen in the languages that have many resources.
In the production of a language resource, many linguists understand that it has already taken time to collect or record the raw data. The next efforts are to analyze what kind of treatments that fit with the data such as finding features to be extracted to useful information. To collect informative features, people need to build language tools and know how to extract those features. Then, to extract those features we also need data that have been annotated. Therefore, we have to decide what annotation schemes are appropriate with the data, that will support our efforts to mine the features, based on the analysis of previous experimental results. After applying proposed methods, we still need to validate whether the annotation really benefits to the data extraction and whether the extracted features appropriate with the goals.

We have worked to develop language resources on the sentences written by L2 learners for Bahasa Indonesia (Irmawati et al., 2016a). We also defined a dependency relation annotation scheme for Bahasa Indonesia (Irmawati et al., 2017a) by considering the language characteristic. Then we, annotated sentences with the annotation and trained the MST Parser to build a dependency relation parser. The parser was used to annotated a large sized of Indonesian corpus to extract their dependency relations. The dependency relation features has been proofed successfully improved a preposition error correction performance (Irmawati et al., 2016b; Irmawati et al., 2017b).

We realized that language learners will use different vocabulary from ones used by the natives. Moreover, the development of languages for people in Indonesia is influenced by their local, indigenous, languages, which may more than one languages. Therefore, the term '*second language learners*' in Indonesia are not only fit with foreigners, but also for the Indonesian people. For that reason, we are currently working to generate specific word lists based on an age level (e.g. children and teenagers). To find the specific words for each age level, we calculate a word that has higher occurrence in a document but lower occurrence in the documents of different age level. Later, we will build a game for children so they may develop a simple sentence using the word list from their age level (easy words). The game may also be used by language learners in beginner level.

Aside of those resources, we also implemented author-

*Corresponding author.
**This work was done when the author was a student in Institut Teknologi Sepuluh Nopember, Indonesia.

topic modelling to map scientific papers with prospective committee members who better review them (Pradina and Khairunnisa, 2018). We also built a mobile application to write a word in Balinese script. User types the word in alphabet then the application will convert it to a unicode so the related word will be printed in *Bali Simbar* fonts (Aranta et al., 2018).

2. Language Resources

These works resulted parallel learner sentences (sentences written by L2 learner with their correction sentences) and a dependency relation annotation scheme.

2.1. Language Learner (L2) Resources

L2 learner resources contained mistakes made by second language (L2) learners that were collected from mistaken sentences written by the L2 learners taken from lang-8 website[1]. The mistaken sentences have their corrections (corrected by their native speakers) so the pairs are considered as parallel data (one contained mistakes and one is its correction).

The size of the original mistaken sentences were limited. It is only 6,559 learners' sentences with the vocabulary sized of 8,673 words. We firstly experimented on preposition errors. To increase its size, we produced artificial error sentences from formal sentences that were taken from Leipzig corpus (Quasthoff et al., 2006). We proposed two methods to inject the correct sentences. The first method, *embeddings*, used dependency-based word embeddings to find a normal sentence that its verb and its preposition object were similar to the verb and the preposition object of the learner sentence that has incorrect preposition. Then, we replaced the preposition in the normal sentence with the incorrect preposition taken from the learners' sentence (Irmawati et al., 2016b). The second method, *selection*, selected randomly injected sentences in which the mistakes highly resembled the mistaken sentences written by L2 learners (Irmawati et al., 2017b).

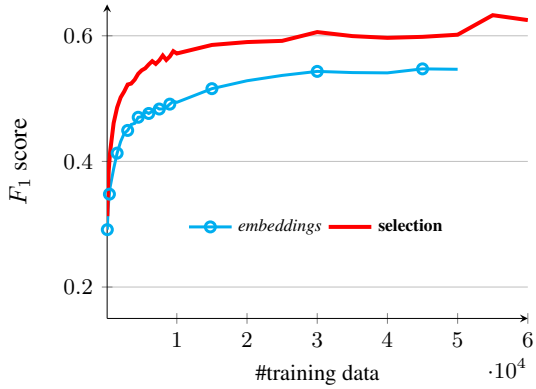


Figure 1: The *selection* method presents better results than the *embeddings* method.

We evaluated the resulted artificial sentences by using them as a training data to correct the preposition errors. The *embeddings* method was promising because it took lesser time to generate artificial L2 learners' sentences. However, our experiments resulted that the *selection* method performed better as showed in Figure 1. The results also shows that training using more data resulted better but the best results were obtained by the data resemble to the data contained mistake. Both experiments used dependency features, so we may conclude that the dependency relation annotation scheme really benefits to preposition error correction for Indonesian.

2.2. Dependency Relation Annotation Scheme

As explained in Subsection 2.1., the verb related to a preposition is important as well as the object of the preposition. Therefore, dependency relation annotation scheme is important to train a dependency relation parser. We defined a dependency annotation scheme (Irmawati et al., 2017a) and annotated 1,132 sentences to build the dependency parser model with accuracy of 81.2% for UAS.

Our work is an adaptation of the Stanford annotation scheme for English proposed by de Marneffe and Manning (2013). This annotation scheme was really useful to our task in Section 2.1. because without that dependency relation features, we obtained less than 50% F_1-score in the preliminary experiments.

2.3. Word List Based on Age Level

In the context of language learning, word choices are critical and be a foolishness for learners. Many researchers tried to help them by developing a simple word list (Coster and Kauchak, 2011) to help learners understand documents not in their spoken language, in the side of vocabulary and structure.

Our goal is to generate vocabularies used by children and teenagers and to justify whether the two different age levels use similar vocabulary in Indonesian. We took data by crowdsourcing from the internet (`https://hai.grid.id/` and `https://bobo.grid.id/`). As the comparison document, we also crowdsourced a national newspaper (`detik.com`) to comply that the two targeted vocabulary were different from one used in the formal articles.

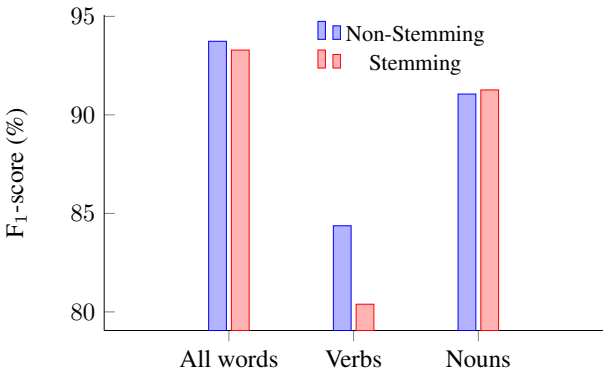


Figure 2: The results of classification on documents consumed by different age levels show that Non-Stemming words performs better.

[1] `http://lang-8.com`

Like other text processing, we also preprocessed the data with tokenization, case folding, stop words filtering, and stemming. We then calculated the TF-IDF for each word related to each document class. Then we used *Multinomial Naïve Bayes* as the classifier. We did some experiments by differentiated the dataset of using all word types, only verbs, and only nouns. For each experiment, we did with stemming and without stemming the words. The experimental results show that all words features performed better than individual word type experiments, which means that the POS of the words did not contribute well.
Figure 2 shows that the dataset of all words and all verbs without stemming performed better. This phenomenon indicates that the existent of affixes influences the complexities of a word. However, the experiment need to be extended to justify what affixes mostly used in the lower age level articles. In Indonesian, every time an affixes is added to a word, it may derived the word to different POS tag. Therefore, it is necessary to evaluate whether several times word derivations may increase word complexity.
The experiment also listed some words contained in specific class as shown in Table 1. However, we have not find a specific pattern that may be concluded. Therefore, we need to work further to find some interesting pattern from the data.

Affixes Words	**Categories**		
	Children	**Teenager**	**Newspaper**
antara	antaranya	antaranya	di antara, antaranya
buruk	buruk	buruknya	memperburuk, memburuk
budaya	kebudayaan kebudayaan-nya	berbudaya	budayanya,
negara	negara	senegara-nya	kenegaraan

Table 1: Some affixes words based on age level category

2.4. Assigning Prospective Reviewers to Scientific Papers

As a conference attracts high number of participants, it would be time consuming for program committee to assign reviewers to bunch of submitted papers. Moreover, they need very wide knowledge to pair varied paper topics and reviewers' expertise to make sure that each paper will obtain the best valuable comments.
We employee *author-topic modelling* to solve this situation. The method is very simple. We assumed that authors who wrote a published paper are ones with expertise-related to the topic of the paper. Then, we trained a model for those relations. Next time a program committee need to assign reviewers to a submitted paper, they may apply the trained model to the submitted paper to find the best reviewers who match with the paper.
In this experiment, we collected 422 scientific papers from various conferences in Indonesian, from 2013 to 2016, from `http://is.its.ac.id/pubs/oajis/`. We did three scenarios, said *without-stemming*, *with-stemming*, and *only nouns*. The experimental results showed that the *with-stemming* scenario got the lowest perplexity and the highest topic coherence with 100 number of topics as shown in Table 2. The *with-stemming* scenario even obtained -1.528 for the 50 number of topics.

	Without Stemming	**With Stemming**	**Only Nouns**
Number of topics	50	100	150
Mean of perplexity	171.18	127.46	85.17
Std. Dev. of perplexity	1.158	1.040	0.667
Topic coherence	-1.765	-1.615	-1.741

Table 2: The results of three scenarios

We analyzed that a problem that may affect the results are the number of papers written by an authors, which is very few during a year. It is usually only about two to three papers when he/she wrote as the first author. Moreover, there are possibility of an author to write little bit different topic from his/her main area though the area may have some similarities in context.

3. A Transliteration for Balinese

As other languages in Asia, some parts in Indonesia used scripts to write their ancient documents. The script contains some rules because one letter represents one syllable with inherent vowel /a/. Each letter covers the consonant, vowel, and some accent speech. The languages are still be used in the daily conversation with some simplification in vocabularies and language levels. It also differentiates the vocabulary spoken to elderly and honour people.
On the other hand, to face of the diversity, Indonesian tend to use their national language that is used as a unity language. The use of alphabet to write the documents also supports the development of the national language. Therefore, the local language speakers are decreasing. In the spirit of preserving endanger languages in Indonesia, we developed an application as an input method for Balinese script (Aranta et al., 2018). Balinese is a language used in Bali, Northern Nusa Penida, Western Lombok and Eastern Java. It is a *Malayo-Polynesian* language spoken by 3.3 million people (as of 2000). Balinese is not mutually intelligible with Indonesian. Some words in the higher level are almost similar to Javanese but ones used in the daily conversation have different in meaning[2]. In the case of Balinese, the script can be written with 18 consonants and 9 vowels[3]. The rules are *Gantugan*, *Gempelan*, *PasangPageh*, numbers, and punctuations.
We used the official *Balinese dictionary* to generate rules and to represent letters in unicodes, known as *Bali Simbar*. Then we used test data taken from Balinese Galang Foundation, a preservation institution for Balinese culture.

[2] `https://en.wikipedia.org/wiki/Balinese_language`
[3] `https://en.wikipedia.org/wiki/Balinese_script`

The evaluation was done on 151 letters containing 13 letter types taken from Balinese dictionary. It obtained 92.72 % accuracy that means that there are some unrecognized letters including ones related to pronunciation. The difficulties to represent correct letter were because not all users knew well how to pronounce a letter. For example, they cannot differentiate 'e' and '$\bar{e}$' in the word 'p$\bar{e}$k$\bar{e}$n'.

4. Conclusion

We described the process of collecting a learner corpus and how to generate data artificially to extend the original learner corpus. We have tried two artificial generation data methods involving dependency-based word embeddings and selection methods. By increasing its size, we showed that the preposition error correction system developed from the data, resulted better performance. Moreover, we also found that the dependency relation features also improved the performance.

For the document classification based on age level, we concluded that the document classifications works better by involving all words without stemming. However, there is still remaining work to confirm whether the complexity of affixes might be used to identify whether a word is difficult for a low level age such as children. In our experiments, we list some words that only appear in one age level. Therefore our work to build a word list of vocabulary used in an age level may still find other possible results. On the other hand, our experiments on author-topic model gave different results as the stemming scenario performed the best.

Our work on the transliteration of Balinese will also be continued to reverse the process from the script to alphabet letter for easy readability. Also it may be continue to be applied to other scripts used in local, indigenous, especially in the west part of Lesser Sunda.

5. Acknowledgements

The **L2 language resources works** were supported by *Directorate General of Higher Education* through DIKTI scholarship, Indonesia and the Computational Linguistics Laboratory of Nara Institute of Science and Technology (NAIST), Japan.

The **Balinese transliteration** was partially supported by the Indonesian Ministry of Research, Technology and Higher Education grant number 113/UN48.15/LT/2018.

6. Bibliographical References

Aranta, A., Gunadi, I. G. A., and Indrawan, G. (2018). Utilization of Hexadecimal numbers in Optimization of Baliness Transliteration String Replacement Method. In *Proceedings of the 11th AUN/SEED-Net Regional Conference on Computer and Information Engineering*, pages 3746–3753, Surabaya, Indonesia, Nov. IEEE.

Coster, W. and Kauchak, D. (2011). Simple English Wikipedia: A new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 665–669, Portland, Oregon, USA, June. Association for Computational Linguistics.

de Marneffe, M. and Manning, C. D. (2013). *Stanford Typed Dependencies Manual: Revised for Stanford Parser v.3.3 in December 2013*. September. Revised for the Stanford Parser v.3.3 in December 2013.

Irmawati, B., Komachi, M., and Matsumoto, Y. (2016a). Towards Construction of an Error-Corrected Corpus of Indonesian Second Language Learners. In Francisco Alonso Almeida, et al., editors, *Input a Word, Analyse the World: Selected Approaches to Corpus Linguistics*, chapter 27, pages 425–443. Cambridge Scholars Publishing, Newcastle upon Tyne.

Irmawati, B., Shindo, H., and Matsumoto, Y. (2016b). Exploiting Syntactic Similarities for Preposition Error Corrections on Indonesian Sentences Written by Second Language Learner. In Sakriani Sakti, et al., editors, *SLTU-2016, 5th Workshop on Spoken Language Technologies for Under-resourced Languages*, volume 81 of *Procedia Computer Science*, pages 214–220, Yogyakarta, Indonesia. Elsevier.

Irmawati, B., Shindo, H., and Matsumoto, Y. (2017a). A dependency Annotation Scheme to Extract Syntactic Features in Indonesian Sentences. *International Journal of Technology (IJTech)*, 8(5):549–558, November.

Irmawati, B., Shindo, H., and Matsumoto, Y. (2017b). Generating Artificial Error Data for Indonesian Preposition Error Correction. *International Journal of Technology (IJTech)*, 8(3):957–967, April.

Pradina, R. and Khairunnisa, S. O. (2018). Author-Topic Modelling for Reviewer Assignment of Scientific Papers in Bahasa Indonesia. In *Proceedings of 2018 International Conference on Asian Language Processing (IALP)*, pages 351–356, 11.

Quasthoff, U., Richter, M., and Biemann, C. (2006). Corpus Portal for Search in Monolingual Corpora. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 1799–1802, Genoa, Italy.

Proceedings, pages 235–238
Paris, UNESCO Headquarters, 5-6 December, 2019.

Languages and Technology in Bhutan

Tshewang Norbu, Tenzin Namgyel
Secretary, Sr. ICTO
Dzongkha Development Commission, Thimphu, Bhutan
tshewangnorbu@yahoo.com, tenraj.1047@gmail.com
{tnorbu, tnamgyel}@dzongkha.gov.bt

Abstract

Bhutan, a small country with population less than a million, is linguistically rich with 19 different spoken languages. Dzongkha is the national language and also the official language of Bhutan. Dzongkha Development Commission is a government institute mandated to formulate language plans and policies, develop and promote the national language, and research, document, preserve and protect other indigenous languages of Bhutan. This paper attempts to provide an overview of the languages of Bhutan, language policies and plans, current status of technology development for Dzongkha, and the challenges being faced by Bhutan

Keywords: Dzongkha, Bhutan, Dzongkha Development Commission, Language Technology

Résumé

མི་སྒོ་བས་ཁྱི་གཅིག་ལས་ཉུང་བའི་འབྲུག་གི་རྒྱལ་ཁབ་ཆུང་ཀུ་ནད་ ཁ་སྐད་མ་འདྲཝ་ ༡༩ དེ་ཅིག་ཡོད་པ་ལས་ཁ་སྐད་ཀྱི་ཕྱུག་པའི་རྒྱལ་ཁབ་ཅིག་ཨིན། རྫོང་ཁ་འདི་འབྲུག་གི་རྒྱལ་ཡོངས་སྐད་ཡིག་ཨིན་པའི་ཁར་ གཞུང་འཛིལ་གྱི་ཁ་སྐད་ཡང་ཨིན། རྫོང་ཁ་གོང་འཕེལ་ལྷན་ཚོགས་འདི་ སྐད་ཡིག་གི་སྲིད་བྱུས་དང་འཆར་གཞི་བརྩམ་ནི་དང་། རྒྱལ་ཡོངས་སྐད་ཡིག་གོང་འཕེལ་དང་དར་ཁྱབ་གཅདཔ་ནི། དེ་ལས་ལུས་པའི་ཁ་སྐད་ཚུ་ ཞིབ་འཚོལ་དང་ཐོ་བཀོད་འབད་དེ་ཉམས་སྲུང་དང་བདག་འཛིན་འཐབ་ནིའི་འགན་དབང་ཡོད་པའི་ གཞུང་གི་གཞུག་སྟེ་མཐོ་ཤོས་ཅིག་ཨིན། ཡིག་ཆ་འདི་ནང་ འབྲུག་གི་ཁ་སྐད་ཚུ་དང་། སྐད་ཡིག་གི་སྲིད་བྱུས་དང་འཆར་གཞི། རྫོང་ཁའི་དོན་ལུ་འཕྲུལ་རིག་གོང་འཕེལ། དེ་ལས་གདོང་ལེན་ཚུ་གི་སྐོར་ལས་ངོ་སྤྲོད་འབད་ནི་ཨིན།

1. Introduction

Guided by a unique development philosophy, Gross National Happiness, Bhutan, a tiny country sandwiched between China in north and India in south places so much importance to her rich cultural heritage and linguistic diversity. The land area of Bhutan is 38,394 square km and population is 734,374 (as of 2018).

1.1 Languages of Bhutan

According to the official survey carried out in 1991, there are 19 different spoken languages. The latest edition of Ethnologue have listed 23 languages. The list excludes Tibetan and includes two foreign languages, namely Kurux and Hindi, and Nupbikha, Lunanakha and Layakha as three additional languages of Bhutan. The Bhutanese languages are classified under Central Bodhish, East Bodhish, Bodic, and Indo-Aryan (van Driem 1998)

Dzongkha is the national language of Bhutan. It was declared as the national language in 1971. It is the native language of eight of the twenty districts of Bhutan, viz. Thimphu, Pünakha, Paro,Wangdi Phodrang, Gasa, Haa, Dagana and Chukha in western Bhutan, but Dzongkha is spoken as a lingua franca throughout the country.

According to Pema Wangdi (2015), all the languages of Bhutan with the exception of Dzongkha, Tshangla, and Lhotsham(Nepali), fall under the category of “endangered” languages. Three languages, namely Monkha, Lhokpu, and Gongduk are critically endangered. One dialect kown as Olekha, a variety of Monkha spoken in Rukha under Wangdue Dzongkhag, is a moribund.

1.2 Dzongkha Development Commission

The Fourth King Jigme Singye Wangchuck established Dzongkha Development Commission (DDC) in 1986. It is now a premier government agency with the highest authority in the matters related to Languages. The broad mandates of the DDC are to formulate language plans and policies, to carry out the activities to develop and promote Dzongkha, as the national language and to carry out the activities to preserve and protect other indigenous languages of Bhutan as the rich linguistic and cultural heritage of Bhutan

The overall plans and policies of the DDC are guided by the commission which consists 10 members. The chairperson of the commission is the Honorable Prime Minister of Bhutan.

2. Writing Systems

The script used to write Dzongkha is the same script used for Tibetan. This writing system consists of 30 consonant symbols and 4 vowel symbols. It is called the Uchen Script and it is one of the two scripts first developed by Thonmi Sambhota in the 7th century, the other being Ume. Uchen is based on the Devanagari script. Another type of script known as Joyig, which is unique to Bhutan, was first developed by Demang Tsemang in Bhutan in the 8th century. Joyig has exactly the same number of consonant and vowel symbols and those Joyig symbols represent the same phonemes as the Uchen script.

'Ucän	ཀ་	ཁ་	ག་	ང་	ཅ་	ཆ་	ཇ་	ཉ་	ཏ་	ཐ་	ད་	ན་	པ་	ཕ་	བ་
Jôyi	ཀྱ་	ཁྱ་	གྱ་	ང་	ཅ་	ཆ་	ཇ་	ཉ་	ཏ་	ཐ་	ད་	ན་	པ་	ཕ་	བ་
Roman	ka	kha	ga	nga	ca	cha	ja	nya	ta	tha	da	na	pa	pha	Ba

'Ucän	མ་	ཙ་	ཚ་	ཛ་	ཝ་	ཞ་	ཟ་	འ་	ཡ་	ར་	ལ་	ཤ་	ས་	ཧ་	ཨ་
Jôyi	མ་	ཙ་	ཚ་	ཛ་	ཝ་	ཞ་	ཟ་	འ་	ཡ་	ར་	ལ་	ཤ་	ས་	ཧ་	ཨ་
Roman	ma	tsa	tsha	dza	wa	zha	za	ha	ya	ra	la	sha	sa	ha	a

Figure 1: 30 Consonants of Dzongkha

The four vowel symbols in Dzongkha are ི(གི་གུ་g'ikhu) representing the sound "i", ུ(ཞབས་ཀྱུ་zh'apju) representing the sound "u", ེ(དགྲེད་མོ་drengbo) representing the sound "e" and ོ(ན་རོ་naro) representing the sound "o". When no vowel is indicated, the vowel in a syllable is automatically "a". Dzongkha language is syllabic. A Dzongkha syllable can have 1 to 7 characters and most interestingly, up to four characters can stack on one another as shown in the figure below.

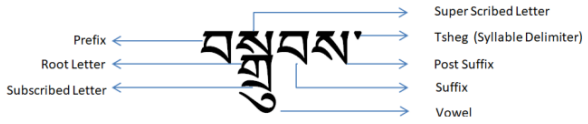


Figure 2: A Dzongkha Syllable

Since the number and the type of phonemic inventories of the Bhutanese languages are quite similar, albeit some with few extra ones, this script can be used to write Dzongkha and other languages of Bhutan with almost the same degree of consistency and economy. Dzongkha is the only officially written language of Bhutan. Other native languages are written sporadically for interpersonal communication in informal settings and in social media. Lhotshamkha is written using the Nepalese script (Nepal Bhasa), one of the Brahmic scripts of the Kathmandu valley of Nepal.

3. Language Policies, Plans and Actions

The language policy framework for Bhutan is enshrined in its Constitution, the mother of all laws. Section 8 of Article 1 of the Constitution states, "Dzongkha is the national language of Bhutan"; and, in Section 1 of Article 4, "language" and "literature" are enumerated along with the other cultural heritage of Bhutan to be preserved, protected and promoted. It provides a clear framework for developing and promoting Dzongkha as the national language of Bhutan and it also provides a clear framework for preserving the other languages of Bhutan as the rich linguistic and cultural heritage of Bhutan.

Pema Wangdi (2015), summarizes the language policy of Bhutan in the form of Quadrilingual Model:

Dzongkha, Chökê, English, and Mother Tongue

Mother Tongue in this model is the sum total of all the mother tongues in Bhutan.

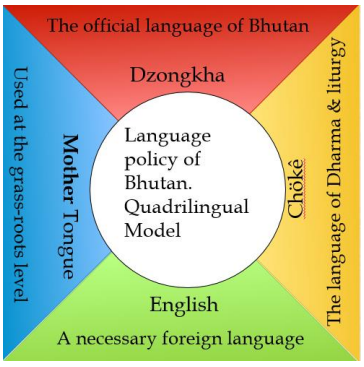


Figure 3: Quadrilingual Model

Dzongkha efficiently serves as the official language of Bhutan while Chökê serves as the language of Dharma and liturgy; and English is apparently used as the necessary foreign language while mother tongues are used at the grass-roots level.

Today, under the framework of 12th Five Year Plan, DDC is working to achieve following five objectives:

1. Develop and promote Dzongkha.
   To standardize the spelling and grammar, DDC writes grammars and dictionaries and other reference books. Various promotional activities such as competitions are also being carried out.

2. Research and document indigenous languages.
   Writing trilingual lexicons and phrase books between Dzongkha, English and indigenous languages and carrying out preliminary researches on the grammar of indigenous languages are some of the important activities carried out to preserve our indigenous languages.

3. Enhance Dzongkha usage in public service delivery.
   Despite Dzongkha being the national language, wide usage of English to disseminate information and deliver services to the general public gives unequal access to information and services and causes inequalities in the society. To tackle this and to protect our national identity, DDC provides support in making websites, applications and forms accessible in Dzongkha.

4. Enhance usage of Dzongkha in education system.
   Role of national language and other indigenous languages in mainstream education is limited as English is the only medium of instruction and Dzongkha was taught only as a language subject. DDC recommends and assists in including more Dzongkha subjects and also actively participates in curriculum development.

5. Develop language technology and promote Dzongkha through ICT.
   To keep Dzongkha digitally alive in the world of technology, DDC develops Unicode compatible Dzongkha keyboards and fonts. To bridge the digital

divide, solve other inequalities caused by the language barrier, and protect our national language, DDC has started to work on developing technology for Dzongkha which is covered in more detail in the following section.

4. Language Technology Development

In Bhutan, English is used very widely with technology and this can lead to digital extinction of national language and other language, and cause real extinction gradualy. Moreover, language barrier is the cause of unequal access to knowledge, information, services and digital divide which creates inequalities in the society. The best and the only solution would be to develop technology for Dzongkha.

4.1 Encoding, Input and Rendering Supports

Twenty years ago, there was no recognized or de-facto standard for encoding Dzongkha or Tibetan script characters. Word-processing applications and other programs adapted for Dzongkha used a variety of ad-hoc non-standardized encodings which gave codes in character sets actually meant for encoding Roman characters to Dzongkha letters. The biggest obstacle in using electronic Dzongkha data was the fact that files could not be easily shared by different Dzongkha word-processing programs and other applications without converting files from one encoding scheme to another.

In 2000, a 3-year project was carried out to develop a standardized system for Dzongkha based on the new Unicode / ISO 10646 character encoding standard. During the project, a standard keyboard layout was developed, a locale for Dzongkha and collation rules were developed and Unicode compatible fonts were also developed. Input and rendering support for Linux operating system was developed by Department of Information Technology and Telecom (DITT), Ministry of Information and Communication (MOIC) under the framework of PAN localization project phase I (2004-2007).

Today, major operating systems like Windows, MacOS iOS and Android have built-in Dzongkha rendering and input supports. However, as of now, except for older version of Linux, no language packs or localized versions of operating systems are available in Dzongkha.

4.2 Automation and Processing of Language

Currently, technology for Dzongkha is limited to input, storage and display. There are no working Dzongkha language processing tools: not even spelling or grammar checker. Except for few research works carried out by DITT, MOIC during the 2nd phase of PAN localization project (2007-2012), there is no history of much work done in the field of natural language processing in the past.

The DDC, in collaboration with College of Science and Technology, Phuntsholing, Bhutan started to develop part of speech (POS) tagged corpus since 2014. With support from Indian Institute of Technology (IIT), Guwahati, India, we have achieved the following.

4.2.1 Dzongkha Word Segmentation

Dzongkha does not have any word delimiter like space in English. Therefore, it is necessary for the computer to do word segmentation to be able to progress further in Dzongkha text processing. The Dzongkha word segmentation is done as a syllable tagging problem using various NLP toolkit and the best model is 95% accurate as of now.

4.2.2 Dzongkha Part of Speech (POS) Tagging

DDC has developed Dzongkha POS tagset and also annotated POS to around 2 lakhs Dzongkha words, manually. Using the manually POS-tagged corpus, we trained a model which is around 90% accurate as of now. We expect the performance to improve with increase in corpus size.

4.2.3 Dzongkha Automatic Speech Recognition

The first Dzongkha automatic speech recognition (ASR) prototype was developed during the ASR summer school conducted by IIT, Guwahati in the year 2017. With the knowldege and motivation gained, DDC has now increased the speech corpus to around 15 hours of recording and we are hoping to increase it.

5. Challanges

Bhutan as a least developed country, has many other priority areas to invest on for wellbeing of the citizen. Though Bhutan fully understands the importance of developing technology for language, not much could be done so far without enough fund. We also don't have required expertise in this field. For instance, we do not have anybody with master's degree or PhD. artificial intelligence; forget about in computational linguistic or natural language processing specialist. With the limited state funding, DDC is unable to provide long term training to staff and we are also unable to participate in international forums. Bhutan has not been able to secure any funding support from outside for development and promotion of language technology.

6. Conclusion

For a small landlocked country like Bhutan, cultural heritage and linguistic diversity is very important for national identity. Preservation of rich cultural heritage is one of the pillars of our development philosophy of Gross National Happiness. A sound language policy that ensures protection of all our languages is enshrined in our mother of law, the Constitution of Bhutan. To keep our national language and other languages alive, it is important that we make them usable in the digital world but due to the lack of expertise and fund, we could do very little so far to process by computer. Technology for our national language is limited to input, storage and display. We solicit support in

terms of expertise, capacity building and funding from international organizations and donors.

7. Acknowledgements

We would like to thank Indian Institute of Technology, Guwahati, Assam, India for helping us in testing our corpus on various NLP toolkits and developing models for word segmentation, POS-tagging and ASR.

8. Bibliographical References

Pema, W., Language Policy and Planning in Bhutan, available at https://www.dzongkha.gov.bt/uploads/files/articles/A_Paper_on_Language_Policy_&_Planning_in_Bhutan_by_Pema_Wangdi_c8e8caeee831129a3be15aa6e99732c2.pdf accessed on 25th Dec, 2019

Chungku, C., Jurmey, R.,Gertrud, F. 2010. "Building NLP resources for Dzongkha: A Tagset and A Tagged Corpus," Proceedings of the 8th Workshop on Asian Language Resources, pages 103–110, Beijing, China, 21-22 August 2010, pp. 103-110.

Department of Information Technology & Telecommunications (DITT), Bhutan, Research papers, Available online at "https://www.dit.gov.bt/research-paper," Accessed on 25th Dec 2019.

Norbu, S., Choejey, P., Dendup, T., Hussain, S. and Muaz, A., 2010. "Dzongkha Word Segmentation", Proceedings of the 8th Workshop on Asian Language Resources, COLING2010, Beijing, China, April 3-8, pp. 200-209.

Proceedings, pages 239–242
Paris, UNESCO Headquarters, 5-6 December, 2019.

Towards Speech Technologies for Romani Language in Slovakia

Milan Rusko[1], Sakhia Darjaa[1], Marián Trnka[1], Róbert Sabo[1], Štefan Beňuš[1,2]
Institute of Informatics of the Slovak Academy of Sciences,
Dúbravská cesta 9, 845 07 Bratislava, Slovakia
[2]University of Constantine, the Philosopher in Nitra,
Štefánikova 64, 949 07 Nitra, Slovakia
{milan.rusko, sakhia.darjaa, marian.trnka, robert.sabo, stefan.benus}@sav.sk

Abstract

This work summarizes activities at the Institute of Informatics of the Slovak Academy of Sciences (IISAS) in the research and development of speech technologies in the language of the Roma minority in Slovakia. Basic facts on orthography, phonetics, and prosody of the Romani language are given. The design of the text corpus, the speech database and speech synthesizers is described. The challenges that still have to be tackled in speech recognition are briefly mentioned. The current research in human-machine communication using robotic head is presented and its possible use in the education of Romani children is discussed.

Keywords: underresourced languages, Romani language, speech technologies, Romani speech synthesis, robotic head.

Résumé

Táto práca sumarizuje doterajšie aktivity Ústavu informatiky Slovenskej akadémie vied vo výskume a vývoji rečových technológií v rómčine, popisuje súčasný výskum a plány do budúcnosti. Uvádzajú sa základné fakty o pravopise, fonetike a prozódii rómskeho jazyka. Popisuje návrh textového korpusu, rečovej databázy a rečových syntetizátorov. V krátkosti sa spomínajú výzvy, ktoré je ešte potrebné vyriešiť v rozpoznávaní reči. Prezentuje sa samotná práca na komunikácii človek-stroj pomocou robotickej hlavy a diskutuje sa o jej možnom využití vo výučbe rómskych detí.

Motto: "Sa tu man šaj kames, kana dumakeri chib nadžanes? Sar saj prindžarav tiro, jilo kana tiro lav hin gadžikano?" (How could you want me if you do not know my language? How could I know your heart, if I do not understand the Gadžo (non-Roma) words?) – fragment of a poem by J. Berky-Luborecký**.**

1. Anglelav – Introduction

Digitally endangered languages are languages used by people who are too few in number or too poor to make them attractive to commercial software developers. This means that native speakers of these languages end up having two barriers to overcome to access computers – first, they have to learn English; then they have to learn IT skills. Their native language is marginalized, and becomes digitally endangered. (www.mealldubh.org, 2008) From this point of view the language of Romanies in Slovakia (Romani chhib) without any doubt belongs to the digitally endangered languages. The total number of Roma in Europe amounts to 6.6 million people. (http://romani.uni-graz.at, 2008) Until recently Romani was only an oral language, without a written norm. During the last decades an attempt to create a written norm started in different countries and in several cases the written form was codified.

1.1 Serviko Romani čhib – The Language of Eastern Slovak Romanies

The language of European Romanies belongs to the group of Indian languages. Their language was always influenced by their habitat, where they stayed as a nomadic nation for certain time. Persian, Greek, Armenian, or even Slavonic words can still be found in this language. Slovak Romanies have in their language many words which were adopted from standard Slovak and even more expressions "borrowed" from local Slovak dialects. The language of Romanies living in the south of Slovakia uses also many expressions of the Hungarian origin.

During the population census in 2001, 89 920 people in Slovakia have claimed that their nationality is Roma. Nevertheless the real number of the members of Roma community is estimated to be as high as 380 000.

In Slovakia – as opposed to the Czech Republic the Romani language (as a group of varieties) does not seem to disappear, although many of the dialects and the local varieties (especially in Western, Central and Southern Slovakia) are endangered or close to extinction. But mainly in Eastern Slovakia in some socially isolated localities a gradual change from traditional bilingualism to monolingualism in Romani – or to a radical lowering of competence in Slovak (in generations which grew up to a productive age after the revolution in 1989) can be observed. This is in relation to the growth of unemployment and disintegration of social nets. (Elšík, V., 2007)

The Eastern Slovak (Serviko) Romani is one of the three main dialects that are spoken by Romanies in Slovakia. It is spoken by approximately 80–85% of Roma population in Slovakia, therefore Eastern Slovak Romani dialect was chosen as the basis for grammar, lexicon and phraseology of the codified language.

The term Servika comes from the words "Serbika", "Serbos", "Serbija", reflecting the fact that they came from Serbia. Text.

1.2 Irišagos, Pheniben, Vakeribno Melodija – Pronunciation, Orthography, Intonation

In Slovakia the Romani orthography was codified in 1971 and recodified in the nineties. The codified form is based on the orthographical rules of Slovak. Basically the Slovak alphabet was adopted including diacritical marks, which are however used according to different rules. The most recent publication available on the topic (Hübschmannová, M., et al. 2006) gives an overview on the grammar and phonetic rules of the Romani language. In Romani a phoneme is always written with the same corresponding grapheme. This rule rule is consistently followed with the following phonemes: a, b, c, č, čh, d, e, f, g, h, i, j, k, kh, ľ, m, n, o, p, ph, r, s, š, t, th, u, v, z, ž. Unlike in the Slovak language, palatalized d,t,n,l are written with hacek-accents even before vowels i and e.

There is a general rule is in Romani, that voicing and aspiration is neutralized at the end of words. In the written form however the graphemes corresponding to the aspirated and voiced phonemes are preserved (jakh [jak] – jakha [jakha] (eye-eyes)).

Combinations of vowels in the foreign words (in Slovak ia, ie, iu) are written as ija, ije, iju (geografija, gimnazijum). In contrast to Slovk, in which both /i/ and /y/ graphemes refer to the same [i] vowel, Romani does not use y grapheme.

The phoneme set used in codified Serviko Romani is very similar to that of standard Slovak. The only Romani specific phonemes that do not exist in Slovak are ?h, kh, ph and th, which are pronounced with a slight aspiration. Without the aspiration the words have essentially different meaning, e.g.: pherel (draw, pump) – perel (fall), khoro (jar) – koro (blind)(Hübschmannová, M., et al. 2006). Hence, aspirated sounds are separate phonemes in Serviko Romani.

A problematic issue, significant also for prosody is the problem of vowel quantity. Long vowels are not marked by acute accent.

The rule that voicing is lost at the end of words holds also for Slovak and therefore does not cause any difference in pronunciation between the languages.

Word stress is generally placed on the pre-final syllable in Serviko Romani, which also holds for Eastern Slovak dialects. However, too many exceptions exist from this basic rule, so we had to use a lexicon of word stress exceptions. (As our rule-based intonation model was built for Standard Slovak, which has accents always on the first syllable, we had to revise the whole model and change the rules of accentuation. The rules for phoneme lengths prediction, which are based mainly on the mean value of the phoneme length remained unchanged.).

2. Chibakro modulatoris – Speech synthesizer

Since the beginning of the millennium several versions of Romani synthesizers have been developed at the Institute of Informatics. Technically they can be considered as four generations of synthesizers: Diphone synthesizer, Unit-selection synthesizer, HMM (Hidden Markov Model) synthesizer and DNN (Deep Neural Network) synthesizer. We will briefly describe each version.

2.1 Diphone Synthesizer

The first Romani speech synthesizer was based on concatenation of recorded realizations of diphones, and it was a slightly modified version of our Slovak synthesizer. To prepare a diphone database we had to define a set of words that contain the Romani specific diphones that were not present in the diphone set of the Slovak speech synthesizer. For the baseline version adding a set of diphones containing aspired phonemes – ph, kh, ch, th was sufficient. Table 1 presents examples of Romani words containing various vowel - aspired consonant combinations.

čh / $tʃ^h$	lačharela ačhavel čhamenger	e čhercheňa bilače prečhinel	vičhinel dičhiben	očhohano ľočhipena	čhuvaleskro oduchárel
kh / k^h	khamoro naarakheha	te khelel jekhetane	dikhipena lokhiben	o khosno polokhe	te khuvel mukhavkerel
ph / p^h	phabaj zaphenel	phenel barephikeskero	phirel priphandel	phosavel dophenel	phuv phurikano
th / t^h	tharel sathemeskro	themeskero prethovel	ithiskero prithovibe	te thovel odothar	thudeskero thuvalel

Table 1: Examples of Romani words containing aspirated consonants.

Grapheme to phoneme conversion was based on a sophisticated set of rules supplemented by a pronunciation vocabulary and a list of exceptions. The Romani version of this block was created by changing Slovak rules according to Romani pronunciation rules mentioned in section 1.2.

With a relatively small intervention in the text preprocessing, grapheme to phoneme rules, phoneme inventory and the intonation model of the Slovak speech synthesizer we managed to create a basic level diphone speech synthesizer in Serviko Romani. The quality of the speech was adequate to the type of the synthesizer. The segmental quality was not perfect, but was acceptable. In spite of the shortcommings – robotic and buzzy speech quality and very simple prosody model – the produced speech was reasonably intelligible.

2.2 Unit-Selection Synthesizer

The Unit-selection Romani synthesizer was using our own synthesis engine presented in Rusko, Trnka and Darjaa (2006). The algorithm did not calculate the joint and cost functions, but was merely relying on phonetical–phonological pre-selection of elements (mainly syllables). The main features determining the selection were phonological context, pitch and phoneme length. The unit-selection synthesizer with a CART trees (Breiman et.al., 1984) based prosody model brought much better naturalness and allowed for more advanced experiments in prosody modeling.

2.3 Statistical parametric HMM synthesizer

The synthesizer using Hidden Markov Models was developed using the HTS toolkit (Zen et.al., 2007). It was designed in order to be able to generate not only emotionally neutral speech, but also warning messages in Romani. This feature was meant to be used for automatic generation of voice messages in the warning information system in case of fire, flood, state security threats, or other crisis situations.

2.3.1 Text Resources

To get a better idea of the structure of the language and to prepare a set of sentences for the speech database recording, a bigger volume of texts was needed. For the Romani language, the quantity and quality of the texts available is highly insufficient. Moreover, many of the texts that have been published are written in the particular local dialect of their author and not in the standardized form of the language.

To have at least some amount of texts for initial efforts, we used the archive of the Slovak Romani newspaper "Romano nevo l'il" from the years 2003 to 2010, unpublished Romani fairy tales by Vladimír Zeman and several tenths of pages of texts that we were provided by Stanislav Cina, who is our Romani language expert and experienced bi-lingual voice talent. We obtained only about 600 kB of texts in total from these sources, which formed our corpus of Romani texts. These were analyzed and a basic set of 1574 phonetically rich sentences was selected for the recording of the emotionally neutral part of the Romani speech database. The same amount of Slovak texts was prepared and recorded for training of the Slovak baseline neutral voice.

2.3.2 The Expressive Speech Synthesizer

In 2012 we designed an expressive speech database CRISIS. It consists of 90 prompted short warning messages (160 sentences) per level of urgency and per language. An original three-step method of recording this expressive speech database was proposed and successfully employed. The sentences were uttered by one male bilingual speaker in three levels of urgency. The first one represents neutral speech and served mainly as a reference level to the higher two levels. The second level represents assertive warnings or commands, and in the third level the speaker uttered the messages in extremely intense and urgent way – "as if human lives were directly endangered and the speaker had to try to save them" (Rusko et.al., 2012).

As it was mentioned in the previous paragraph, a larger neutral speech databases were recorded by the same speaker in both languages to create higher quality neutral baseline voices, that can be later adapted to the three final voices with different levels of expressivity.

The HTS system (Zen et.al., 2007) was used for creating the speech synthesizer. The baseline HMMTTS voice was trained from the emotionally neutral bigger speech database in the corresponding language. This voice was then adapted to three levels of expressivity using the recordings of the emotional speech database CRISIS and applying the Constrained Structural Maximum A-Posteriori Linear Regression (CSMAPLR) technique (Nakano et.al. 2006).

According to the informal listening tests the synthesized speech kept the voice quality, rhythm, intonation, and the resulting expressive load from the source recordings very well. Different levels of urgency of the messages were reliably distinguishable across the three adapted synthesizers (level 1 – normal/neutral, 2 - urgent, and 3 – extremely urgent) in both Slovak and Romani languages. The results suggest that the used "three step method" of expressive speech database development is suitable for gathering a good quality expressive and hyper-expressive speech database for the design of speech synthesizers for emergency situations.

2.4 Statistical parametric DNN synthesizer

The Slovak Deep Neural Network (DNN) synthesizer was developed using the Merlin toolkit for building DNN models for statistical parametric speech synthesis (Zhizheng, Watts, and King, 2016). We combined it with our own front-end text processor and the WORLD vocoder (Morise, Yokomori, and Ozawa, 2016). WORLD vocoder decomposes the input speech into three parameters: fundamental frequency (f0), spectral envelope and aperiodicity (representation of excitation via the band-aperiodicity function). The used DNN has six feed-forward hidden layers having 1024 hyperbolic tangent units each.

The DNN synthesis based on the WORLD vocoder is considerably more natural than the HMM-based synthesis we used before.

3. Towards speech recognition in Romani

The issues making the development of automatic speech recognition in Romani are the same as for the other under-resourced languages. The general issues are lack of written texts, lack of speech recordings, lack of speakers suitable for studio recording, lack of annotators knowing the language, and lack of funding. The specific problem is, that the official codified language is spoken only by several tenths to hundreds of Roma people in Slovakia. All other speak their local dialect that can be significantly different from the codified one.

The experiments with Romani speech recognition are ongoing. They are trying to take the advantage of the fact that high quality Slovak acoustical models have already been developed and are trying to overcome the problems with missing data on the Romani-specific triphones and other phenomena. Due to the very limited amount of text data, building a more general state-of-the-art language model is practically impossible. Therefore, the designers can only work on applications that would do with grammars or simple language models.

For the initial experiments we use the ASR system based on Kaldi Speech Recognition Toolkit (Povey et al., 2011). We hope our first results will be ready for publication soon.

4. Teaching L2 and L1 communication skills with a robotic head

A Furhat is a physical 3D humanoid head that employs the optical projection of an animated facial model (Al Moubayed et. al, 2012). The face projections has functionality allowing for eye brow movement, blinking, and various emotional expressions that are easy to adjust or scale.

In our current work (Beňuš, Sabo, and Trnka, 2019) we investigate how the social robotic head Furhat might be used in human-machine communication research.

We designed a novel communicative game "Guess the animal" to study various parameters of human speech, dialogue phenomena, as well as the effectiveness and convenience of the communication. About 100 healthy and 10 handicapped subjects played the game; they were recorded and filled a questionnaire.

Figure 1: A handicapped person playing the Guess the animal with Furhat.

Both healthy subjects and handicapped people would certainly not prefer Furhat over a human. Nevertheless, they expressed a strong positive evaluation on the usefulness of the robot in training and teaching communicative skills. So the social robotic head can probably also be used to assist human teachers in improving skills in second language acquisition with healthy students and with people of various communication handicaps. Moreover, we think this method could be successfully used to raise interest in learning and training the communicative skills also in the codified version of the language of Roma minority. This could help the children speaking different dialects of Romani to acquire and accept the codified form of their language.

5. Conclusion

However the aim of the work was not to create a perfect speech synthesizer, recognizer, or social assistent but to study the under-resourced language, find its main peculiarities, and prepare the basic speech processing background needed for further development of much more comprehensive speech technology applications like pedagogical tools and information systems in Romani.

6. Acknowledgements

This work was supported by VEGA grant nr. 2/0161/18. The authors have included in this work some parts of texts of their publications (Rusko et.al., 2006, 2008, 2012), and (Beňuš, Š., Sabo, R., and Trnka, M., 2019). This was necessary for giving a consistent picture of various phases of the research and providing basic facts on the Romani language that have already been published earlier.

7. Bibliographical References

Al Moubayed, S., et. al, (2012), Furhat: A Back-Projected Human-Like Robot Head for Multiparty Human-Machine Interaction, *in Esposito A., et.al. (eds) Cognitive Behavioural Systems. LNCS, vol 7403*. Springer, Berlin, Heidelberg.

Beňuš, Š., Sabo, R., and Trnka, M., 2019, Word guessing game with a social robotic head, (19th Conference Information Technologies: Applications and Theory, ITAT 2019), *in: CEUR Workshop Proceedings: Iinformation technologies - application and theory 2019, 2019, vol. 2473*, pp. 1-5.

Breiman, L. et.al., (1984) *Classification and Regression Trees*. Chapman Hall, New York.

Elšík, V. (2007), *personal communication*.

http://www.mealldubh.org/index.php/2006/02/05/strength-inconfederation/ (2008)

http://romani.uni-graz.at/rombase/index.html (2008)

Hübschmannová, M., et al. (2006). *Rules of Romani Orthography* (in Slovak). State Paedagogical Institute, Bratislava.

Morise, M., Yokomori, F., and Ozawa, K., (2016), WORLD: a vocoder-based high-quality speech synthesis system for real-time applications, *IEICE transactions on information and systems*, vol. E99-D, no. 7, pp. 1877-1884.

Nakano, Y., et.al. (2006), Constrained Structural Maximum A Posteriori Linear Regression for Average-Voice-Based Speech Synthesis, *Proc. of ICSLP 2006.*

Povey et al., D., (2011), The Kaldi Speech Recognition Toolkit, *in: Proceedings of ASRU 2011.*

Rusko, M., Trnka, M., and Darjaa, S. (2006). Three Generations of Speech Synthesis Systems in Slovakia. *In: Proceedings of XI International Conference Speech and Computer, SPECOM 2006.* Sankt Peterburg.

Rusko, M., et.al., (2008) Making Speech Technologies Available in (Serviko) Romani Language. *In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2008. LNCS (LNAI), vol. 5246*, Springer, Heidelberg, pp. 501–508.

Rusko, M., et.al., (2012), Expressive speech synthesis database for emergent messages and warnings generation in critical situations. *In: Language Resources for Public Security Workshop (LRPS 2012*), *LREC 2012 Proceedings*., Istambul, pp. 50–53.

Zen, H., et.al., (2007) The HMM-based speech synthesis system version 2.0. *In: Proc. of ISCA SSW6*, Bonn.

Zhizheng, W., Watts, O., and King, S., (2016), Merlin: An Open Source Neural Network Speech Synthesis System, *Procedings of 9th ISCA Speech Synthesis Workshop (SSW9)*, September 2016, Sunnyvale, CA.

Proceedings, pages 243–246
Paris, UNESCO Headquarters, 5-6 December, 2019.

Enabling Linguistics Diversity and Multilingualism Worldwide

NTeALan - AI/NLP/NLU Platforms For Sharing and Leveraging African Language Resources For Education In Africa

Elvis MBONING[1], Jean-Marc BASSAHAK[1], Juanita Fopa[1], Jules ASSOUMOU[2], Damien NOUVEL[3]
NTeALan
Makepe, parcours vita - Douala, Cameroon
{levismboning, bassahak, juanita.fopa}@ntealan.org[1]
julesassoumou@yahoo.fr[2], damien.nouvel@inalco.fr[3]

Abstract

Among the challenges, the African continent faces the issue of safeguarding and enhancing its cultural and linguistic inheritance. Created in 2017, the NTeALan project aims at setting up intelligent tools for the digitization, development and teaching of African languages. Since 2019, our association is supported by 20 volunteers from various fields, several partnerships were signed (universities, institutions, startups, association) and we have implemented several prototypical applications: chatbots, collaborative dictionary, linguistic map, REST APIs for African language resources. With its multimedia center dedicated to NLP, NTeALan aims at making African languages a cornerstone of Africa's cultural, linguistics and technological development.

Keywords: African languages, AI, NLP, NLU, Lexicography, Education, Resources, API, Culture

Résumé (Bassa'a language spoken in Cameroon)

i kété mítìík mí mám má má nlámá hólós áfríkà í màŋgéŋ máná dì gwě, màhɔ̀l má má ntágɓɛ́nɛ́ í máhɔ́p més nì bíɓɔ̀ŋɔ́l gwés,íɲǔ hàlà ɲɛ́n ntealan tòhálá kìí̀ à má sál ŋgándàk mú í ndzɛ́l ĩ, à ŋgí sálák ní láná lɛ́ɛ́: à ńtí bínóŋɔ́l bí nɔ́ndɔ́ bí bí ńhólá lɛ̀ɛ̀ dí nìí́gá, nì hólɔ̄s màhɔ́p més lóŋní ndzì bìɓɔ̀ŋɔ́l bí mɔ̀ndɔ́. náànɔ́ í ŋwìí dìkóó díbàà nì mbɔ̀gí dzóm nì bɔ̀ó, í sí màtíŋmá ≪núlóŋú ntealan≫,í lɛ́l móó màà (mɔ̀ɔ màà) má ɓót má lólàk í mítèn mí bíɓɔ̀lɔ́ ŋwɔ́mísó, ɓá ńsál ntóŋwádá nì ɓès íɲǔ màhɔ̀l má lítúŋgá lí áfríkà. àndàk màtìŋmá yé ǹsángɛ̀, mápɯmá kɛ̀ɓá nsàngɛ̀ lóŋnì mìntén mí bíɓóló ŋgàndàk: bísùkúlù nì mínlóŋbìkŋí. ŋgàndàk í màm í ɛ̀ɓóŋà, í màm má gáhólá lítùŋgá lí áfríkà dzɔ́ lísó, díúlà símá: api íɲùú téèdà mítén mí bíɓóŋɔ́l gwés, bíkààt bí bíɓúk, bíkààt bí máhɔ́p…). ndàp í ɓóló í nlp/nlu í yè í hálá (càmàlùn). ntealan à ńsómból ɓɔ̀ŋlɛ̀ɛ̀ màhɔ́p má áfríkà má ɓá ŋŋɛ́m ú máhɔ̀l má bílɔ̀n gwés gwɔ́ bísómá.

1. Overview on NTeALan project

Among the challenges, the African continent faces the issue of safeguarding and enhancing its cultural and linguistic inheritance. Created in 2017[1], and managed by academics and the African Learned Society, NTeALan (New Technologies for African Languages) is an Association that works for the implementation of intelligent technological tools for the promotion, development and teaching of African national languages. Our goals are to digitize, safeguard and promote African national languages through digital tools and Artificial Intelligence to build a new generation of young Africans aware of the challenges of appropriating the languages and cultures of the continent. In this paper, we want to present our first major activities realised between 2017 and 2019 with the NTeALan's teams. We will continue with the difficulties encountered and the future challenges for the upcoming years.

1.1. Why is this project a necessity for Africa ?

Language plays an important role in defining the identity and humanity of individuals. As Tunde Opeibi (Tunde, 2012, p.272) said "In Africa, evidence shows that language has become a very strong factor for ethno-national identity, with the ethnic loyalty overriding the national interest". To date, the African continent has more than 3000 languages, more than two thirds of which are poorly endowed. Among the reasons justifying this observation, we can list:

- The lack of a strong linguistic policy in favor of these languages,
- The lack of specialists in NLP / AI / NLU trained on the continent and specialists in these languages,
- The lack of linguistic resources (textual and oral mainly oral tradition) and NLP/NLU tools available for most of these languages,
- A virtual absence of these languages in the digital space (social networks, online platform, etc.) and in the educational system.
- No African scientific community dedicated to technological issues related to the written and oral transmission of knowledge in African languages
- Few standardized African languages on a vast ensemble and their gradual disappearance over the years

[1]Mainly by Elvis MBONING (NLP Research Engineer at INALCO) and Jean Marc BASSAHAK (Contractor, Web designer and developer). Jules Assoumou, Head of the Department of Linguistics and African Literature at University of douala, joined us later.

Faced with this, specialists from various fields (computer developer, NLP engineer, academics specialized in questions of linguistics, cultures, didactic and African pedagogy) gathered around the world to create an association which was to make it possible to set up NLP and NLU tools based on the current state of the art of AI in order to create collaborative environments for the creation of sharing resources and tools around African languages intended to serve the scientific community, companies, social networks and all other public or private institutions.
We are convinced, as Tunde Opeibi (Tunde, 2012, p.289) already said so well that "the linguistic diversity in Africa can still become the catalyst that will promote cultural, socio-economic, political, and technological development, as well as sustainable growth and good governance in Africa."

1.2. Our strategies

Our approach is exclusively based on the collaboration model (Holtzblatt and Beyer, 2017). We want to allow African people to contribute to the development of their own mother tongue, supervised by specialists and academics of African languages. Our model involves setting up several communities: the community of speakers of these languages, the community of native specialists (guarantors of traditional, cultural and linguistic knowledge), the community of academics specialized in African linguistics technologies and the community of social / institutional / public partners. The chart 1 below summarizes this strategy.

Collaboration	Research and Tools
Create a community of professionals specialized in sociolinguistics and technological issues (NLP / NLU / IA) for Africa	Create open source platform for academics specialists in other to give them more means for their research activities
Create a community of volunteer contributors around collaborative platforms for building common language resources (text, image, video)	Systematically equip all the African national languages by language family and encourage their use by young people in Africa
Collaboratively use created resources to set up an autonomous language teaching platform	Help public/private institutions and companies by integrating our technologies in the education system, in their own platforms, and others.

Table 1: Strategy adopted by NTeALan

1.3. The main NTeALan's projects

During our first years[2], we initiated some projects mainly centered on building collaborative, multilingual and distributed resources for African languages and cultures. Our team has just put in place:

- The multilingual conversational agents platform [NTeABot] to teach young African students their mother tongues
- The collaborative dictionaries (lexicon, audio, picture and video) platform for African national languages and cultures [https://ntealan.net/dictionnaires],
- REST and Websocket APIs for sharing African language resources [https://apis.ntealan.net],
- The platform for the management of lexical and terminological resources in African languages [https://ntealan.net/dictionaries-platform],
- The dictionary platform illustration [http://upfiles.ntealan.org/koken],
- The tool for digitizing documents in African languages
- The dictionaries annotation platform (component of the numerisation tool) [http://dico-edit.ntealan.net],
- Scientific research activities in Natural Language Processing (NLP) and Natural Language Understanding (NLU) for African languages,
- The management of the NTeALan center at Makepe (Douala) and many others internally.

Essentially based on REST and Websocket APIs technologies, these first initiated projects are still in the testing phase in a few languages. Indeed, we started from a few pilot languages (essentially Bantu and semi-Bantu languages)[3] already having available resources in low quantity. Our objective was to apply the first versions of our NLP/NLU tools (morphological, syntactical and semantic analysis, NER, automatic conjugation and POS tagging) on this sample in order to analyze the results and see if these could be generalized on others in the same linguistic family. The figure 1 below show the general structure of our actual system.

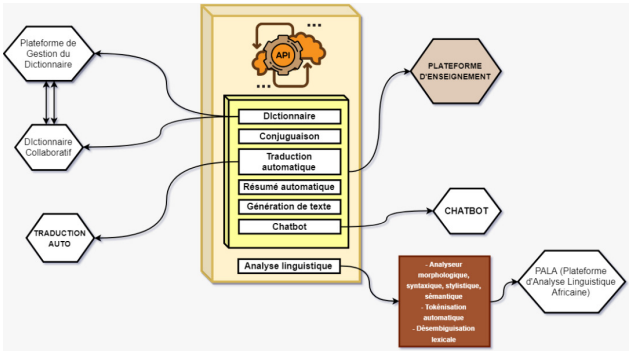


Figure 1: NTeALan APIs and services infrastructures

[2]We started at the end of 2017 till now.

[3]We have yɛmba, bassa'a, duala, ngiembon (spoken in Cameroon) and bambara (spoken in Mali)

2. Description of our current major projects

Our main projects, on which others depend, are the collaborative dictionaries for African languages resources and tools (cf. figure 2), the African linguistic map and the conversational agent platform [NTeABot] for teaching African languages.

2.1. Open source collaborative dictionaries, NLP/NLU tools and their REST API

For this first main project[4], we give access to native speakers and experts who have an expertise in African language to build collaboratively resources like lexicon[5], illustration of cultural phenomenon, sound and videos (recording process) based on semantic information on article in their native language. These shared resources are freely available for all contributors through our REST API hosted at [https://apis.ntealan.net/ntealan/dictionaries].

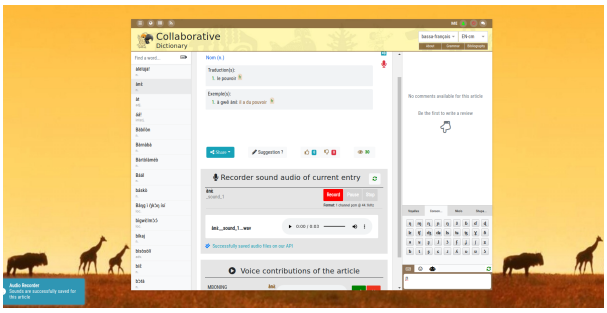


Figure 2: Collaborative dictionaries for sharing multimodal and multilingual resources in African languages. This platform is under license on Creative Commons BY-SA: [https://ntealan.net]

AI being essential today for the construction of good quality linguistic resources and tools, we lead internally with our academic partners (language and African literature department of the University of Douala, the ERTIM team of the INALCO) numerous research activities in Artificial Intelligence, NLP, and NLU in order to contribute to the industrialization of these languages. The results of these studies help our applications and can be use by others researchers: this include data (in different common formats like TEI (Benoit and Turcan, 2006), LMF, XMLAF[6]) and tools.

2.2. Multilingual conversational agent platform and its REST API

For the second main project (cf. figure 3), we want NTeABot platform to teach young African students in Africa or at the diaspora in their different languages in the same time during course teaching with teachers at school and with their parents at home. With NTeABot platform we can also build other competences (applications) such as information on time, definition on wikipedia, informations on NTeALan dictionaries, on NTeALan project and some others. The test version is available on our official website [https://ntealan.org]

Figure 3: Sample of discussion about NTeALan dictionaries with NTeABot agent. Tested on [https://ntealan.org]

2.3. African linguistic and cultural map

We realised a detailed linguistic and cultural map on each country in Africa (cf. figure 4). For each of these countries, we provided a number of spoken languages, dialects, their classification, development status and the place where they are used. We want here to enumerate all the living languages in Africa in other to link them with their resources and NLP/NLU tools. The actual hosted version only contains the Cameroonian languages. We will add for others African countries in our next deliveries next year depending on the availability of language resources.

3. Problems encountered and futures challenges

The implementation of these first projects enabled us to note certain important problems. In upcoming years, it will be a question of filling them up and making them more mature for future deadlines. First of all, let's start with the problems encountered.

3.1. Problems encountered

We are currently facing two main problems in the NTeALan association:

[4]This project was born following the research work of Elvis MBONING at the University of Douala and University of Lille 3 (Master thesis): (MBONING, 2016) and (MBONING, 2017). We can cite other related works in this fields like (Assoumou, 2010), (Mangeot and Enguehard, 2011), (Vydrin et al., 2016), (Maslinsky, 2014), (Nouvel et al., 2016), etc.

[5]For this work, we also build another platform to manage lexicographic resource: [https://ntealan.net/dictionaries-platform].

[6]NTeALan codification format of dictionary for African bantu and semi-bantu languages

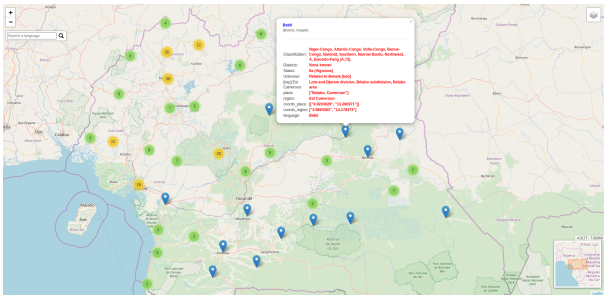


Figure 4: Linguistic description map of African languages. The figure show precisely Cameroon area with focus on bebil language spoken in the city of Belabo at the East region of Cameroon

- the first is the number of our researchers and insufficient IT resources. For the staff, we do not yet have all the specialists (NLP, NLU, African languages) that we wish to treat. Regarding IT resources, we do not have enough robust IT infrastructures (servers, field tools) for our research work on these languages.
- the second is the lack of funding to fill our research and development activities for the NLP and NLU tools concerned. Our funding mainly come from the members of the association, which is insufficient in light of our current ambitions.

3.2. Next challenges to be met

Our ambitions are great and will require more staff and financial resources.

- Above all, we want to make the greatest number of specialists in African languages and cultures in each African country and in the world, to join our association to further challenges.
- Find funding from private and public institutions, businessmen, companies who can support our research work and the continuous development of our applications for the teaching of these poorly endowed languages.
- Improve and enrich all existing platforms and open them up more to the scientific community and speakers of these languages. We will focus on : the autonomous platform for teaching languages and cultures, the conversational Agent Assistant for Language Teaching and Virtual cultural museum for the safeguarding of the African socio-cultural inheritance, etc.
- Strengthen our partnerships with African social and cultural institutions, universities, research laboratories and companies specializing in our research areas. The aim is to create a community of experts in linguistics, technological and cultural issues across the continent

4. Conclusion

All in all, NTeALan is: 2 years of existence (2017-2019), 3 institutional and private partners, more than 5 students trained in computational linguistics, 1 equipped research and development center based in Douala (Cameroon), more than 15 members worldwide, 3 tooled african languages, more than 10 internal projects and 5 open source and collaborative resources and tools. We are motivated to continue to grow up and to build an active environment around data, researchers, african people and technological companies for these languages. You are free to support or join us: contact@ntealan.org | dons@ntealan.org

5. Acknowledgements

This project is actually supported by the ministry of post and telecommunication of Cameroon, Department of linguistics and African literature of the University of Douala (Cameroon), Research teams ERTIM of INALCO (France) and Fractals system (France). We can also cite: Daniel BALEBA, Juanita FOPA, Merci Christian BONOG BILAP, Marcel TOMI BANOU, NTOMB David, NTOMB Nicolas, Théophile KENGNE, Parfait DIMONO, Yves Bertrand DISSAKE, Daniel NOUVEL and many other contributors.

6. References

Assoumou, J. (2010). *Enseignement oral des langues et cultures africaines à l'école primaire*. Éditions Clé, Yaoundé, Cameroun, 1st edition.

Benoit, J.-L. and Turcan, I. (2006). La TEI au service de la transmission documentaire ou de la valorisation des richesses patrimoniales : le cas difficile des dictionnaires anciens.

Holtzblatt, K. and Beyer, H. (2017). 7 - Building Experience Models. In Karen Holtzblatt et al., editors, *Contextual Design (Second Edition)*, Interactive Technologies, pages 147–206. Morgan Kaufmann, Boston, January.

Mangeot, M. and Enguehard, C. (2011). Informatisation de dictionnaires langues africaines-français. In *journées LTT 2011*, page 11.

Maslinsky, K. (2014). *Daba: a model and tools for Manding corpora*.

MBONING, E. (2016). De l'analyse du dictionnaire yémba-français à la conception de sa DTD et de sa réédition sur support numérique. Mémoire Master 1, Université de Lille 3.

MBONING, E. (2017). Vers une métalexicographie outillée : conception d'un outil pour le métalexicographe et applica- tion aux dictionnaires Larousse de 1856 à 1966. Mémoire Master 2, Université de Lille 3.

Nouvel, D., Donandt, K., Auffret, D., Maslinsky, K., Chiarcos, C., and Vydrin, V. (2016). Resources and Experiments for a Bambara POS Tagger. *Intra Speech*, page 14.

Tunde, O. (2012). Investigating the Language Situation in Africa. In *Language and Law*, Language rights, pages 272–293. Oxford Handbooks in Linguistics, Great Clarendon street.

Vydrin, V., Rovenchak, A., and Maslinsky, K. (2016). Maninka Reference Corpus: A Presentation. In *TALAf 2016 : Traitement automatique des langues africaines (écrit et parole). Atelier JEP-TALN-RECITAL 2016 - Paris le*, Paris, France, July.

Proceedings, pages 247–250
Paris, UNESCO Headquarters, 5-6 December, 2019.

Multilingual Neural Machine Translation in Low Resource Settings

Pulkit Madaan, Fatiha Sadat
IIIT Delhi, UQAM
Delhi India, Montreal Canada
pulkit16257@iiitd.ac.in, sadat.fatiha@uqam.ca

Abstract

Neural Machine Translations (NMT) models are capable of translating a single bilingual pair and require a new model for each new language pair. Multilingual Neural Machine Translation (M-NMT) models are capable of translating multiple language pairs, even pairs which it hasn't seen before in training. Availability of parallel sentences is a known problem in machine translation. M-NMT model leverages information from all the languages to improve itself and performs better. We propose a data augmentation technique that further improves this model profoundly. The technique helps achieve a jump of more than 15 points in BLEU score from the Multilingual NMT Model. A BLEU score of 36.2 was achieved for Sindhi–English translation, which is higher than any score on the leaderboard of the LoResMT SharedTask at MT Summit 2019, which provided the data for the experiments.

Keywords: Neural Machine Translation, Low Resource Languages, Multilingual Transformer, Deep Learning, End-to-end Learning

Résumé

न्यूरल मशीन ट्रांसलेशन (NMT) मॉडल एकल द्विभाषी जोड़ी का अनुवाद करने में सक्षम हैं, लेकिन प्रत्येक नई भाषा जोड़ी के लिए एक नए मॉडल की आवश्यकता होती है। मल्टीलिंग्वल न्यूरल मशीन ट्रांसलेशन मॉडल (M-NMT) मॉडल कई भाषा जोड़े, यहां तक कि वो जोड़े जो प्रशिक्षण में पहले नहीं देखे गए हैं, उनका अनुवाद करने में भी सक्षम है। समानांतर वाक्यों की दुर्लभता मशीन अनुवाद में एक ज्ञात समस्या है। M-NMT मॉडल खुद को बेहतर बनाने और बेहतर प्रदर्शन करने के लिए सभी भाषाओं से जानकारी प्राप्त करता है। हम प्रस्तुत शोध पत्र में एक डेटा वृद्धि तकनीक को प्रस्तावित करते हैं जो इस मॉडल के प्रदर्शन में बड़े दर्जे का सुधार करती है। हमारी नई तकनीक, M-NMT मॉडल से, BLEU स्कोर में 15 से अधिक अंकों की छलांग हासिल करने में मदद करती है। सिंधी-अंग्रेजी अनुवाद के लिए 36.2 का BLEU स्कोर हासिल किया गया, जो MT Summit 2019 में LoResMT SharedTask के लीडरबोर्ड पर किसी भी स्कोर से अधिक है। LoResMT SharedTask के आयोजकों द्वारा प्रदान किया डेटा, इस शोध पत्र में किये गए प्रयोगों के लिए इस्तेमाल हुआ है।

1. Introduction

A lot of the models for end-to-end NMT are trained for single language pairs. Google's Multilingual NMT (Johnson et al., 2017) is a single model capable of translating to and from many languages. The model is fed a token identifying a target language uniquely along with the source language sentence. This allows the model to translate between pairs for which the model hasn't seen parallel data, essentially zero-shot translations. The model is also able to improve upon individual translation qualities too by the help of other languages. We propose to improve this quality further with a data-augmentation technique that's able to improve the BLEU scores two fold. The technique is simple and can work with any model. We show that increasing the amount of data available for training artificially with our technique in a way as simple as just swapping the source with target sentences and using the same sentence as source and target can improve the BLEU scores significantly. Also, we show that since all language pairs share the same encoder and the same decoder, in a case of transfer learning, the model is able to leverage data from rich resource language pairs for learning better translations for low-resource pairs. Using Hindi–English data in training improved the BLEU scores for {Bhojpuri, Sindhi, Magadhi}<>English. The structure of the present paper is described as follows: Section 2 presents the state of the art. Section 3 presents our proposed methodology. Section 4 describes the corpora used in this research. In section 5, we put forward our experiments and evaluations, perform an ablative analysis and compare our system's performance with other Google's Neural Machine Translation(Johnson et al., 2017). Section 6, compares our results with other methods that participated in the LoResMT Shared Task at the MT Summit 2019. Finally in Section 7, we state our conclusions and perspectives for future research.

2. Related Work

Significant progress has been made in end-to-end NMT Cho et al.(2014); Sutskeveret al.(2014); Bahdanau et al.(2015); and some work has been done to adapt it to a multilingual setting. But, before the mulitilingual approach of Johnson et al., 2017, none of the approaches have a single model capable of dealing with multiple language pairs in a many-to-many setting. Dong et al.(2015) use different decoders and attention layers for different target languages. Firat et al.(2016) use a shared attention layer but an encoder per source language and a decoder per target language. Lee et al.(2017) use a single model with the whole model shared across all pairs but it can only be used for a single target language. The model proposed by Johnson et al.(2017) has a single model for a many-to-many task and is able to perform in zero-shot setting too but translating sentences between pairs whose parallel data wasn't seen by the model during training. Arivazhagan

et al.(2019) also propose a model for zero-shot translation that improves upon Google's Multilingual NMT Model (Johnson et al., 2017) and achieves results on par with pivoting. They propose English as the pivot language and use the give the target language token to the decoder instead of the encoder. In order to improve the independence of encoder on source language the similarity between all sentence vectors and their English parallel sentence embedding is maximised along with the minimisation of the translation cross-entropy loss. They use a discriminator and train the encoder adversarially for similarity maximisation. Artetxe et al.(2018) and Yang et al.(2018) also train the encoder adversarially to learn a shared latent space.

3. The Proposed Methodology

The technique we propose is simple consists of four components named **Forward**, **Backward**, **Self** and **High**. **Forward** augmentation is the given data itself. **Backward** augmentation is generated by switching the source and target label in the **Forward** Data, so the source sentence becomes the target sentence and vice versa in parallel sentence pair. **Self** augmentation is generated by using only the required language from the parallel sentences and cloning them as their own target sentences, so the source and target sentence are the same. An example of the augmentations is shown in Figure 1

We know that translation models improve with increase in data and since we also have the same encoder for every language, we can use a language pair that is similar to the language pairs of the task and is a high resource pair to further improve the encoder in encoding source independent embeddings, for transfer learning through the Multilingual architecture of Johnson et al.(2017) . So we propose **Multilingual+** which uses the above mentioned three augmentations (Forward, Backward, Self) along with **High** augmentation; **High** augmentation consists of high-resource language pairs, like Hindi–English parallel data, in Forward, Backward and Self augmentations. This helps in improving the translation models of low resource pairs; {Bhojpuri, Sindhi, Magadhi}<>English.

4. Dataset

Parallel data from four different language pairs are used in the experiments. Following are the language pairs along with the number of parallel sentences of each pair:

1. Sindhi–English (29,014)
2. Magadhi–English (3,710)
3. Bhojpuri–English (28,999)
4. Hindi–English (1,561,840)

Data for pairs 1–3 were made available at the Shared Task at MT Summit 2019. While data for pair 4 was obtained from the IIT Bombay English–Hindi Corpus (Kunchukuttan et al., 2018). The Train-Val-Test splits were used as given by the respective data providers.

5. Experiments

We performed experiments on the Multingual+ model and showed how the addition of each of augmentations we proposed improves the performance by an ablative analysis.

The basic architecture is the same as in Johnson et al.(2017) with the difference of data that is fed. The source sentences get a target language token prepended. Joint Byte-Pair Encoding is learnt for subword segmentation (Sennrich et al., 2016) to address the problem of rare words. Following are the augmentations included in Multinlingual+

- **Forward**
  Sindhi-to-English, Bhojpuri-to-English, Magahi-to-English
- **Backward**
  English-to-Sindhi, English-to-Bhojpuri, English-to-Magahi
- **Self**
  Sindhi-to-Sindhi, Bhojpuri-to-Bhojpuri, Magahi-to-Magahi, English-to-English
- **High**
  Hindi-to-English, English-to-Hindi, Hindi-to-Hindi

To understand how each augmentation improves the BLEU score, we create 4 methods:

- **Base**
  This is the standard model as used in (Johnson et al., 2017), hence it uses only **Forward** and forms our baseline.
- **Base + Back**
  We add **Backward** augmentation to the baseline model
- **Base + Back + Self**
  We add **Self** & **Backward** augmentation to the baseline.
- **Multilingual+**
  This uses all the augmentations:**High** along with **Forward**, **Backward** & **Self**.

Parameters and training procedures are set as in Johnson et al.(2017). PyTorch Sequence-to-Sequence library, **fairseq** (Ott et al., 2019), was used to run the experiments.

Table 1 shows that **Multilingual+** consistently outperforms the others. The table also confirms that the more augmentations you add to the Multilingual NMT model (Johnson et al., 2017), the more it improves. Adding **Backward**, then **Self** and then a new language pair improved the results at each level. All the BLEU scores reported, except star (*) marked, are calculated using SacreBLEU (Post, 2018) on the development set provided.

Augment	Source		Target	
Forward	That town is two miles away.	[English]	वह नगर दो मील की दूरी पर है।	[Hindi]
Backward	वह नगर दो मील की दूरी पर है।	[Hindi]	That town is two miles away.	[English]
Self	That town is two miles away.	[English]	That town is two miles away.	[English]
Self	वह नगर दो मील की दूरी पर है।	[Hindi]	वह नगर दो मील की दूरी पर है।	[Hindi]
High	Is everybody busy?	[English]	Tout le monde est-il occupé ?	[French]

Figure 1: An example of different augments. Here the low resource pair of languages is English–Hindi, and the high resource pair language set is English–French

	Sin-to-Eng	Eng-to-Sin	Bho-to-Eng	Eng-to-Bho	Mag-to-Eng	Eng-to-Mag
Base	15.74*	–	6.11*	–	2.46*	–
Base + Back	18.09*	11.38*	5.01*	0.2	2.55*	0.2
Base + Back + Self	30.77*	18.98*	7.38*	0.6	4.61*	1.2
†**Multilingual+**	**36.2**	**28.8**	**15.6**	**3.7**	**13.3**	**3.5**

Table 1: BLEU scores of different language pairs and directions in the different experiments.
*Results on test data evaluated by the Shared Task at MT Summit 2019 committee.
† Not submitted for the SharedTask

6. Comparisons

We compared our results with other models submitted at the LoResMT Shared Task at the MT Summit 2019. The submission to the Shared Task followed a naming convention to distinguish between different types of corpora used, which we will follow too. The different types of corpora and their abbreviations are as follows:

- Only the provided parallel corpora [-a]
- Only the provided parallel and monolingual corpora [-b]
- Any provided corpora, plus publicly available corpora; including e.g. pre-trained word2vec, [-c]
- Any provided corpora, plus any publicly exterior corpora. [-d]

Using these abbreviations the methods were named in the following manner"

<TeamCode>-<Language-and-Direction>-
<MethodName>-<Used-Corpora-Abbreviation>

Our Team Code was L19T3 and we submitted Base (as Method_1), Base+Back (as Method_2) and Base+Back+Self (as Method_3) all under -a category. Multilingual+ was developed later. Table 2 shows the top 3 performers in different translation directions along with Multilingual+.
Multilingual+ is the best performer in Sin-to-Eng and Mag-to-Eng task, second best performer in Eng-to-Sin and Bho-to-Eng tasks. These results show the superiority of our simple approach. Our data augmentation technique is comparable or better than the best of the methods on the leaderboard of the Shard Task.

7. Conclusion and Future Work

We have presented a simple data augmentation technique coupled with a multilingual transformer that gives a jump of 15 points in BLEU score without any new data and 20 points in BLEU score if a rich resource language pair is introduced, over a standard multilingual transformer. It performs at par or better than best models submitted at the Shared Task. This demonstrates that a multilingual transformer is sensitive to the amount of data used and a simple augmentation technique like ours can provide a significant boost in BLEU scores. With back-translation (Sennrich et al., 2016) being established as a successful approach to augment and train a better translation model, it can be coupled with our approach to experiment and analyse the effectiveness of this amalgam.

8. Bibliographical References

Arivazhagan, N., Bapna, A., Firat, O., Aharoni, R., Johnson, M., and Macherey, W. (2019). The missing ingredient in zero-shot neural machine translation. *CoRR*, abs/1903.07091.

Artetxe, M., Labaka, G., Agirre, E., and Cho, K. (2017). Unsupervised neural machine translation. *CoRR*, abs/1710.11041.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Cho, K., van Merrienboer, B., Çaglar Gülçehre, Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *ArXiv*, abs/1406.1078.

Dong, D., Wu, H., He, W., Yu, D., and Wang, H. (2015). Multi-task learning for multiple language translation. In *ACL*.

Rank	Sin-to-Eng		Eng-to-Sin	
1	L19T2-Sin2Eng-Method3-b	31.32	L19T2-Eng2Sin-Method3-b	**37.58**
2	Base+Back+Self	30.77	L19T2-Eng2Sin-Method2-a	25.17
3	L19T5-sin2eng-xform-a	28.85	Base+Back+Self	18.98
	Multilingual+	**36.2**	Multilingual+	28.8

Rank	Bho-to-Eng		Mag-to-Eng	
1	L19T2-Bho2Eng-Method3-b	**17.03**	L19T2-Mag2Eng-Method3-b	9.71
2	L19T5-bho2eng-xform-a	15.19	L19T5-mag2eng-pbmt-a	5.64
3	L19T5-bho2eng-pbmt-a	14.2	Base+Back+Self	4.61
	Multilingual+	15.6	Multilingual+	**13.3**

Table 2: Top 3 performers in LoResMT Shared Task in different translation directions along with Multilingual+

Firat, O., Cho, K., and Bengio, Y. (2016). Multi-way, multilingual neural machine translation with a shared attention mechanism. *ArXiv*, abs/1601.01073.

Johnson, M., Schuster, M., Le, Q. V., Krikuna, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F. B., Wattenberg, M., Corrado, G. S., Hughes, M., and Dean, J. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Kunchukuttan, A., Mehta, P., and Bhattacharyya, P. (2018). The IIT bombay english-hindi parallel corpus. *Language Resources and Evaluation Conference*, 10.

Lee, J. D., Cho, K., and Hofmann, T. (2017). Fully character-level neural machine translation without explicit segmentation. *Transactions of the Association for Computational Linguistics*, 5:365–378.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels, October. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August. Association for Computational Linguistics.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *NIPS*.

Yang, Z., Chen, W., Wang, F., and Xu, B. (2018). Unsupervised neural machine translation with weight sharing. *CoRR*, abs/1804.09057.

Proceedings, pages 251–254
Paris, UNESCO Headquarters, 5-6 December, 2019.

Komi Latin Letters, Degrees of UNICODE Facilitation

Jack Rueter, Larisa Ponomareva
University of Helsinki, Digital Humanities
jack.rueter@helsinki.fi, dojegpl@gmail.com

Abstract

The Komi-Permyak and Komi-Zyrian language forms are two writing traditions of the pluricentric Komi language, whose first debut in as a literary medium stems from the 14th Century. The alphabets and writing systems have changed numerous times over the centuries, and most published media in the two languages are facilitated by UNICODE. There is, however, one specific time period, where the digital accessibility of plentiful publications is hampered by missing representation in UNICODE, and that is the time span 1932–1938; the Komi languages were printed in a Latin script with numerous and special Latin letters. Due to the poor quality of typography in this era, the characters have been determined transitional, and therefore it has been suggested that characters existing in UNICODE, regardless of range, might be used to alleviate the issue of missing representation. Bearing this in mind, the number of required letters dropped from an original estimation of 18 to 8 or perhaps simply combining descenders. "UNICODE Script Ad Hoc" has provided helpful suggestions for providing digital means for the Komi languages, and with UNICODE in place, the next hurdle will be dealing with font issues. **Keywords:** Komi Latin alphabet, UNICODE, proposal of new characters, digital accessibility, minority languages

Дженыта висьталöм

Перем коми кыв да зырянскöй коми кыв – уна быдкодь местаэзын олісь коми кывлöн гижан кыввез. Коми кылын медодззаись литературнöй артыс аркмис XIV-öт векö. Кад сьöрна коми анбуррез да гижан арттэз унаись вежсьывлісö. Унажык пассэз, кöдна пантасьлöны кык коми кыв вылын лэдзöм литератураын, эмöсь ЮНИКОД-ын. Дзир öтік кадö, 1932-öт восянь 1938-öт воöдз, коми кыввез вылын лэдзöм литература сьöкыта шедö тöдмавны компьютерöн, сідз кыдз сія кадся комиöн гижömись не быд пас эм ЮНИКОД-ын. Эта кадö коми кыввез вылын литература вöлі лэдзöм латинскöй шыпассэзöн, кöдна коласын вöлі уна быдкодь ассяма латинскöй шыпас. Сы кадся типографика эз вöв бур. Эта увья эна пассэз вылö пондісö видзöтны, кыдз вуджан кадся пассэз вылö. Этасянь вöлі шуöм, что пассэз, кöдна пантасьлöны ЮНИКОД-ын быдкодь анбуррезын, вермасö босьтсьыны тырмытöм пас мыччалöм понда. Эта увья колана шыпассэзлöн лыдыс чиніс медодззаись висьталöм 18-сянь 10-öдз. Эта вермис лоны сідзжö пассэзлöн öтлаöтömсянь. "UNICODE Script Ad Hoc" («Юникод Пассэзын торья пантасьöммез») чукöр висьталіс коланаторрез коми кыввез компьютерöн тöдмалöм понда. ЮНИКОД-ö колана пассэз пыртöм бöрсянь пыкöт лоас шрифттэзын, шыпассэзлöн неöтнёжа гижömын.

1. Background

In 2011, a "Language Programme" was drafted at the Kone Foundation in Helsinki, Finland with the objective of promoting research work on endangered languages through funding of individual research projects and the preservation of irreplaceable data sets i.a. In 2012–2013, a Digitization Pilot Project of Kindred Languages[1] was coordinated by Jussi-Pekka Hakkarainen in the auspices of the National Library of Finland to save 1920 and 1930 publications from possible water damage. In the pilot, the National Library of Finland in collaboration with the National Library of Russian in St. Petersburg and with funding from the Kone Foundation "Language Programme" digitized newspapers and school books representative of Balto-Finnic, Mordvin and Mari minority languages. The materials were made available in the Fenno-Ugrica collection at the National Library of Finland[2]

In 2014–2016, the Digitization Project of Kindred Languages[3] was extended to address publications especially representative of the time span 1932–1937 for the Permic, Ob-Ugric and Samoyedic languages.

Due to the express time span 1932–1937, it soon became apparent that little of the Komi publications could, in fact, be totally digitized for UNICODE-based access. Some of the letters were entirely missing. In fact, Komi was not the only one lacking a complete alphabet, but this was also the situation for some of the other minorities of the north, who had a special Latin alphabet developed for them called the Unified Northern Alphabet(Siegel and Rießler, 2015).

Naturally, encoding is not the only matter to be dealt with when digitizing text materials. The characters encoded in UNICODE may also require special glyphs for a given language, which makes it possible to capture a typographically uniform set of data. And it might be argued that language specific word lists and morphology could have an effect on recognition (Silfverberg and Rueter, 2014; Partanen and Rießler, 2019).

[1] https://www.doria.fi/bitstream/handle/10024/94581/Sukukielten%20digitointiprojekti_loppuraportti.pdf?sequence=2&isAllowed=y

[2] https://fennougrica.kansalliskirjasto.fi

[3] https://www.doria.fi/bitstream/handle/10024/130799/Sukukieltendigitointiprojekti_Kansalliskirjasto_Hakkarainen_FINAL.pdf?sequence=2&isAllowed=y

2. Komi and UNICODE

Komi is a member of the Permic branch of the Uralic language family. It is spoken in the Komi Republic, the Perm Krai as well as parts of western Siberia and the Kola Peninsula. Komi consists of a continuum of dialects represented by two modern literary language traditions: Komi-Permyak and Komi-Zyrian. Although most published materials written in the various Komi alphabets are digitally available through UNICODE, there is a time span (1932–1937), when Latinization co-occurred with prolific publication activities, and access to these texts is hampered in digital spheres by the absence of necessary character encoding.

Since an earlier proposal made for an entire extension block to encode Latin letters used in the Former Soviet Union[4] had not been accepted, it was decided that a new language-specific proposal be made from an entirely descriptive perspective. It was important that the description of the missing letters be concise, so as not to be met with an under-informed evaluation of look-alike characters already present in the Latin range of UNICODE. On such look-alike letter can be in the Cyrillic soft sign <ь>, the Latin tone six <ƅ>, and the Latin letter <b>, presented in Figure 1.

	U+044A	U+044C	U+0185	U+0062	U+042A	U+042C	U+0184	U+0042
TIMES NEW ROMAN	ъ	ь	ƅ	b	Ъ	Ь	Ƅ	B
ARIAL UNICODE	ъ	ь	ƅ	b	Ъ	Ь	Ƅ	B
GENTIUM	ъ	ь	ƅ	b	Ъ	Ь	Ƅ	B
CALIBRI	ъ	ь	ƅ	b	Ъ	Ь	Ƅ	B
COURIER	ъ	ь	ƅ	b	Ъ	Ь	Ƅ	B
LUCIDA GRANDE	ъ	ь	ъ	b	Ъ	Ь	Ъ	B
	hard sign	soft sign	tone 6	b	hard sign	soft sign	tone 6	B

Figure 1: Comparing Cyrillic letters soft sign with Cyrillic letters hard sign and Latin look-alikes

2.1. Komi Alphabets with UNICODE Support

The Komi literary traditions are first attested in the form of Old Permic script in 1372. Old Permic scripts are attributed to Stefan Khrap (Saint Stephen of Perm) (1340-1396). Attestation of their use date into the 17th century.

Figure 2: Old Permic rendered in UNICODE 10350–1037F.

From the time of the original Old Permic scripts in 1372 to the present, The Komi language forms have been written in three different alphabet ranges: 1372–1600s in Old Permic (Figure 2), 1600s–Present Russian-related Cyrillic, 1918–1938 Molodtsov Cyrillic (Figure 3), interim 1932–1937 transitional Latin (Figure 4).

While use of the Molodtsov Alphabet is shown with a maximal span of twenty years, it was not used in all publications for the whole time period. In fact, publications made in the Permski Krai did not start using the Molodtsov Alphabet until 1921, and then they abandoned it almost entirely in 1932 for a Komi Latin alphabet.

Figure 3: Cyrillic supplement for Molodtsov in UNICODE 0500–050F.

2.2. Komi Alphabets without UNICODE Support

The Komi Latin alphabet (Figure 4) (1932–1937) consisted of 36[5] letters. Just like the Molodtsov Alphabet, it was a phonemic system, where each character represented an individual phoneme of the Komi languages.

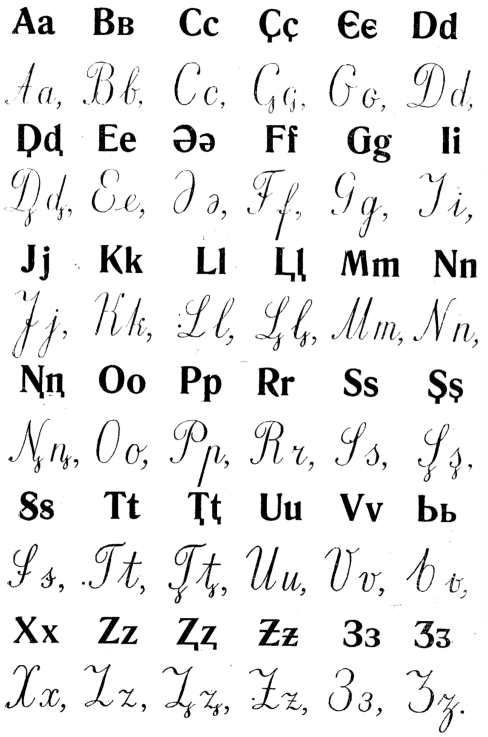


Figure 4: Komi Latin letters A–Z.

The Latin alphabet was used in the Permski Krai from 1932–1937, where it was nearly exclusively used in

[4] http://std.dkuug.dk/jtc1/sc2/wg2/docs/n4162.pdf

[5] The letters Ȝȝ are also rendered as ʒʒ, which increases the number of letters 38.

the publication of both books and newspapers. In the Komi Republic, however, the conversion to the Latin alphabet was never complete. Although it was used systematically in the publication of school books in the Komi Republic (1932–1936), the use of Komi Latin in newspapers varied from mere newspaper names and bolded titles to actual texts. As such falling back on the Molodtsov Alphabet in the Komi Repubic (1936–1938) introduced no radical change for many of the readers.

The Komi Latin Alphabet shares many features of the Unified Northern Alphabet (1931–1937) (Partanen and Rießler, 2019) – and both of these character sets were targeted in a previous proposal to the UNICODE Consortium, mentioned above. The Unified Northern Alphabet, in turn, was an attempt to address phonemic features of under-studied languages of the north, and whose Cyrillification also presented difficulties(Grenoble, 2003).

3. Proposal of New Characters

Originally, it was noted that 18 characters were missing from the Latin Range of UNICODE, and therefor a proposal was made[6] in early June, 2019. No immediate acceptance was expected, but relatively prompt response did open new points of departure. There was something we had overlooked in understanding UNICODE principles.

3.1. Thumb Sketch of UNICODE Principles

A proposal for new characters should address a concrete issue. A given alphabet should derive its letters from a single range. And no precomposed letters should be sought when they can be created utilizing combining diacritics from the (+U0300) section of UNICODE or supplements thereto.

3.2. How We Proceeded

We delimited our proposal to the polycentric Komi language and the digital inaccessibility of over half a decade of literary texts. Here we received support from both the National Library of Finland, Finnish Localization (Kotoistus), and FU-Lab in Syktyvkar, Komi Republic, Russian Federation[7].

We were able to find 27 x 2 letters from within the Latin range, as shown in Table 1.

Aa	Cc	Dd	Ee	Əə	Ff	Gg	Ii	Jj
Kk	Ll	Mm	Nn	Ԉԉ	Oo	Pp	Rr	Ss
~~Ss~~	Tt	Uu	Vv	Xx	Zz	Ҙҙ	Ӟӟ	Ҙ́ҙ́

Table 1: Letters available in Latin Range

These 27 letters, upper and lower case, are readily supported in the Latin range of UNICODE. But this means an addition 10 x 2 letters are required, hence the 18 character proposal.

3.3. After Feedback

Within two months of submission, we received feedback on our proposal. It was noted that the typographical quality of books and newspapers printed in the Komi Latin alphabet were not of high quality – many of the composed letters show symptoms of vaccilation between descender glyphs. It was determined that Komi Latin was, in fact, a transitional alphabet, and therefore it should not be delimited to a single range.

In removing the range requirement, we were immediately given access to 6 x 2 additional letters, see Table 2. It should be noted, however, that these are only typographical solutions, i.e. the Cyrillic Ukrainian IE (+U0404), (+U0454) <*Єє*> is used to represent a non-palatalized voiceless coronal affricate, on the one hand, and Cyrillic VE (+U0412), (+U0432) <*Вв*> is represents a voiced bilabial stop, on the other. Cyrillic letters with descenders are illustrative of glyph differences, which would be addressed in font design for an individual language.

Вв	Ҫҫ	Єє	Ьь	Ҙҙ	Ӟӟ

Table 2: Letters available in Cyrillic Range

The remaining letters (4 x 2) are a set of upper- and lower-case characters that can be enumerated in four alveolars, see Table 3. These lack the same distinctive feature, they are all simply missing a descender.

Dd	Ll	Ss	Tt

Table 3: Letters requiring descenders

To remedy future problems, perhaps a combining descender should also be added to UNICODE.

4. In Conclusion

Drafting a proposal for new characters such as those used in the Komi Latin alphabet (1932–1937), requires good preparation. Previous proposals should be consulted, and all problems presented should be reassessed. Make your arguments precise, and be prepared to accept assessments.

The transitional alphabet scenario is a way to allow for digital accessibility to finite though even extensive materials. For the prolific Komi language materials from the 1930s, this solution is sufficient.

It might be assumed, however, that "transitional alphabet" would be a weak argument for the Cyrillic soft sign look-alike letters <*Ьь*> when dealing with Ingrian (ISO-639-3 izh). This language has only been written in the Latin script.

[6] https://www.unicode.org/L2/L2019/19224-n5101-komi-latin.pdf

[7] The Finno-Ugric Laboratory for Support of the Electronic Representation of Regional Languages https://fu-lab.ru/laboratoriya

All modifications and proposals to UNICODE come at a price, but be receptive and descriptive, and a workable solution is sure to be found.

5. Acknowledgements

Deborah W. Anderson is a member of the ”UNICODE Script Ad Hoc” group. She has provided helpful suggestions for making a successful proposal to UNICODE Technical Committee.
Marina Fedina is the director at FU-Lab in Syktyvkar, Komi Republic, Russian Federation. She has made language materials available to us and introduced us to others involved in the digitization of Komi.
Jussi-Pekka Hakkarainen has provided us not only with access to materials from the Digitization Project of Kindred Languages but with continued advice in our work with various players.
Riitta Koikkalainen is the coordinator of the Finnish Localization Project (Kotoistushanke), and she has been instrumenntal in the drafting of the proposal: ”Komi Latin letters missing in UNICODE”.
Erkki Kolehmainen is a member of the Finnish Localization Project (Kotoistushanke), and he has provided erudite criticism and advice regarding initial drafts of proposal: ”Komi Latin letters missing in UNICODE”.
Enye Lav works at FU-Lab in Syktyvkar, and he has helped locate examples of various glyphs and characters from different years. He has also provided stylized representations of the missing Komi Latin letters as well as an introduction to the entire digitization process at FU-Lab.

6. Bibliographical References

Grenoble, L. A. (2003). *Language policy in the Soviet Union*, volume 3. Springer Science & Business Media.

Partanen, N. and Rießler, M. (2019). An ocr system for the unified northern alphabet. In *Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages*, pages 77–89.

Siegel, F. and Rießler, M. (2015). Uneven steps to literacy. In Janne Saarikivi Heiko F. Marten, Michael Rießler et al., editors, *Cultural and linguistic minorities in the Russian Federation and the European Union*, pages 189–229. Springer, number 13 in Multilingual Education.

Silfverberg, M. and Rueter, J. (2014). Can morphological analyzersimprove the quality of optical character recognition? In *Proceedings of1st International Workshop in Computational Linguistics for Uralic Languages.*, pages 45–56.

Proceedings, pages 255–258
Paris, UNESCO Headquarters, 5-6 December, 2019.

Challenges with Minority Indigenous Languages and Language Technologies

Apolonia Tamata
iTaukei Trust Fund Board
87 Queen Elizabeth Drive, Suva, Fiji
apoloniat@itaukeitrustfund.com.fj

Abstract

In this paper I will describe a few critical features of minority indigenous languages to understand their minority status. The status of standard and non-standard languages and how they co-exist within the same national context will be explained. I will then focus on working with minority indigenous languages and the challenges faced when applying language technologies. The challenges stem from the number and size of the minority indigenous languages. Developing language projects for minority indigenous languages may not be economically efficient for a nation.

Keywords: standard language, non-standard language, multilingual, indigenous minority languages, language technologies

Résumé

E na pepa qo au na dusia tiko e vica na ka era tautauvata kina na veivosanivanua e Viti. Na vosanivanua e sega ni dau taurivaki raraba me vaka na Vosa Vakaviti Raraba. Au na qai vakamacala tale ga e na bolebole e dau sotavi ni tovolei me vakayagataki na porokaramu eso ni kompiuta kei na veilawa. E basika na bolebole ni levu toka na iwiliwili ni vosanivanua ia era lewe lailai na lewenivanua era vakayagataka na nodra vosanivanu. E sega kina ni veiraurau vakailavo na sasaga me toroi cake na vosanivanua ni levu toka na kena iwiliwili e na so na matanitu ia era lewe vica ga na itaukei ni vosa me ra vakayagataka.

1. Minority Languages in Context

A minority indigenous language by implication means that the language is not spoken by the majority of the population of a country. A minority indigenous language (MIL) also implies that other languages are spoken in the country and one of them would be the major language and spoken by the majority. A MIL would therefore not be spoken by the majority of the population. Whether a language is a minority or a majority language in a nation it is relative to the number of speakers there are in each language.

In this context where there are minority and majority languages, it is typical that the population of the country would be bilingual, trilingual or multilingual, that the people may be of different ethnicities or geographical regions and that they would have different mother-tongues. In addition to this, in a multilingual context, there would be an official language and perhaps a national language. The language situation of a country as the above is rich linguistically and culturally and at the same time complex for it denotes language choices, language preferences, language identities and also code mixing.

2. Indigenous Languages in the Pacific

There are approximately 300 indigenous communalects in Fiji (Geraghty, 1983). A communalect is defined as the language the speakers in a village or geographically defined community identify as theirs and is uniform and distinct from the other neighbouring languages (ibid.). Some may refer to these community defined languages as dialects or tribal languages. The principle by which a communalect is defined for the Fijian language recognizes the knowledge of language distinctiveness of the speakers and a sense of identity and ownership towards the communalects.

Apart from communalects, there is the Standard Fijian (SF) dialect of Fiji which does not have a geographically defined community of origin. SF is based on the missionaries' version of a number of dialects and were used initially to write and translate Christian books and teachings. The version was later used for schooling when it started in Fiji. This version became the standard Fijian language as its use extended to religion and also to the early days of education. The new version had prominence and the print form was visible. It had an orthography. It was the language with which to write for the indigenous Fijians. It was and is the language of communication and the media. It is used between speakers of different communalects. It is the other language generally spoken by indigenous Fijians apart from their own community language. The other official languages spoken in Fiji are English and Hindi although the Fiji Hindi which is a creole spoken in Fiji is widely spoken by Indo-Fijians. Fiji's population in 2017 was 884,887[1] however

[1] https://www.fiji.gov.fj/Media-Centre/News/Fiji-Bureau-of-Statistics-Releases-2017-Census-Res

the itaukei or indigenous Fijian population was last indicated in 2012 as 511,838 of the 858,038 which was total population[2]. It can be assumed that the same number or less speak SF. For the purpose of this paper I refer to the 300 Fijian communalects as minority indigenous languages (MIL).

Indigenous languages in Vanuatu, Solomon Islands, New Caledonia and Papua New Guinea are similar to the indigenous Fijian language except that the languages of these countries are languages by linguistic definition and not communalects or dialects where there would be some degree of communicative understanding between neighbouring dialects and communalects. The number of indigenous languages is many, approximately 138 have been recognised in Vanuatu[3], 76 in the Solomon Islands[4], 28[5] in New Caledonia, and 832 living languages of Papua New Guinea[6].

The standard or lingua franca of the people is a creole while the official languages which include the creoles are the international English or French languages or both as in Vanuatu. There is Bislama in Vanuatu, Pijin in Solomon Islands, Tok Pisin and Hiri Motu in Papua New Guinea. The numerous indigenous languages are minority languages as well in their respective countries as the number of speakers are few. The major languages would be the official languages including the creoles.

The languages spoken in the Polynesian and Micronesian islands present different language situations from the Melanesian islands. Per island country, the languages spoken are usually one or two apart from the English or French official language. In these contexts, the indigenous languages would not be minority but majority as most of the population would speak the one indigenous language.

3. Minority Indigenous languages

As described above where there are MIL, there are other languages such as the creoles, the standard and the official languages. The lingua franca have become the standard languages leaving the other indigenous languages as non-standard. Standard languages are then given status as the national language and the official language. They are used in government, education, media and in communication. With these functions they have been developed and have writing systems or the orthography. There are some publications written in them. There are bilingual dictionaries and monolingual dictionaries. These languages have been researched and written about.

The minority languages do not have the same number of speakers or the same level of development as the standard languages. As minority and non-standard, these languages dialects would not have had the same exposure and visibility as standard dialects. Schools and educational curriculum would teach on the standard dialects, the MIL dialects do not share the same recognition. They may exist orally without an alphabet or a writing system. In some cases speakers would use the English alphabet to attempt to write their language. NIL are under developed and under researched although they may be the linguists' delight.

In the language contexts as above the international language takes over the place and space of the indigenous languages in education, media and communication. Standard and international languages further downgrade non-standard dialects however these MIL are the voices and souls of their speakers and are popularly spoken, recited, sung, chanted, hummed and learned as mother-tongues.

4. The Challenges Faced with Minority Indigenous Languages

It is often assumed that the non-standard status of MIL will give way to their becoming extinct as they depend on their speakers to speak them and to transfer language and communication skills to the next generation of speakers. However certain social movements within countries may have given rise to the change in thinking towards language, culture and heritage as vital factors of ethno-identities. People have become more aware and conscious of language and culture and express factors that influence language change and loss. The desire to maintain indigenous languages as forms of identity are expressed in the social media and also in the letters to the editor of local newspapers.

The widely accessible social media forms of Facebook, Twitter, Instagram, Tik Tok and the like have also brought in creativity and innovation in using language to express ideas and to communicate in public. Apart from English, MILs are also used to express in these forums so reports, recounts, feelings, likes and dislikes,

[2]http://prdrse4all.spc.int/system/files/1.2a_census_pop_by_ethnicity_0.pdf

[3] https://www.worldatlas.com/articles/what-languages-are-spoken-in-vanuatu.html

[4] https://www.worldatlas.com/articles/what-languages-are-spoken-in-solomon-islands.html

[5]http://www.world-of-islands.com/Infos/Civilisation/Langues_NC_en.htm

[6] http://valtermoniz.tripod.com/papuanewguinea/id12.html

attitudes and many more can be freely read, interpreted and listened to.

There is creativity in writing and expressing. Phonemes peculiar to MIL are written according or close to the SF writing system and for speech sounds not in SF, the English letters are used or the users would guess the letters to use. Words are shortened, words switch codes from one language to another and standards are ignored. Since MIL are non-standard, the social media has given the speakers avenues for self-expression. As communalects, creoles, national and official languages become visible through social media outlets, conservative language authorities have become concerned about safeguarding indigenous languages.

Similarly, radio programs that broadcast using standard indigenous languages have become creative in their language content with the aim to target the younger generations. The announcers do not limit the language they use nor the content of radio programs to SF as they would create awareness on the other MIL even lesser known pidgins. The radio programs can also be accessed by Fijians living abroad through linking with the internet.

The internet and internet-based technologies have made it possible for languages in particular the MIL to feature outside of their usual geographical locations. However, there are challenges to the making and the maintenance of these language technologies as platforms for language. Having an orthography makes the work easier for both linguists and technologists. A MIL would not necessarily have an established orthography.

Secondly, there may be linguists and there may be technologists but having language technologists are hard to come by. The setting up of websites, language learning technologies, e-books and apps have and should always involve experts from both fields. Needless to say, these experts and the program technologies themselves are quite costly to create as, in many cases, experts would need to be brought in from developed countries. The greatest challenge is faced when deciding whether funds should be spent on MIL with fewer speakers than on standard indigenous languages including the creoles. Either way, maintaining languages through the use of digital technology and the internet is costly. Owners of MIL do not have funds to create and maintain language technologies. In most cases, they would not see the need for making their languages accessible to people outside of their language and culture space using digital technology. Although concerned institutions may provide funds, plans to maintain and monitor language technologies need to be viable and feasible financially.

The resource base of small island countries is limited and has risks. In most cases there would be more pertinent national issues that need relief and funds particularly so now with the impact of climate change and social upheavals creating havoc in communities and national plans. These take priority over MIL initiatives. Another consideration is that intellectual property and copyright approvals on indigenous languages ought to be negotiated prior to implementation because communalects as indigenous languages are regarded by their speakers and owners as cultural property. These measures also rule out misappropriation.

The iTaukei Trust Fund Board (TTFB) is involved with a number of language projects that use digital technology[7]. The two major ones are the SF dictionary App and the Fijian Language and GIS Application. The SF dictionary App is funded by TTFB however the App is now not available on the internet and the overseas based App programmers have severed communications with TTFB. As a highly technical dictionary App both linguistically and technologically, there needs to be more work implemented to keep the project relevant and interactive. In this case the return on funding has not been satisfactory.

The second project investigates the mapping of communalects using GIS. It is titled 'Fijian language and GIS Project'. The project aims to create a language atlas of the Fijian communalects using the geographic information system (GIS), and the second is to develop methods to relate linguistic data to non-linguistic factors that have had influences on people's language use which resulted in language change. The project is spearheaded and principally funded by the National Museum of Ethnology in Osaka, Japan[8]. Specific experts have been engaged and are currently developing the interface. It is also intended that a web-based interface will be developed which aims to show the relationship of linguistic with non-linguistic factors on communalects.

Other language projects that require the use of audio-visual technologies undertaken by TTFB include storytelling and comic animation, and drama performance and video productions. While costs are a major consideration and narrows the scope of projects, the completed products have

[7] http://www.itaukeitrustfund.com.fj/

8 Kikusawa, R. Principle Investigator (National Museum of Ethnology, Osaka), Paul Geraghty (University of the South Pacific), Hirofumi Teramura (National Museum of Ethnology, Osaka), Susumu Okamota (Tokyo University of Foreign Studies), Fumiya Sano (University of Kyoto), John Lowry (Massey University, New Zealand) and Apolonia Tamata (iTaukei Trust Fund Board).

been fruitful. The drama texts and productions are currently being used as literature resources in schools. However, the aim to portray folk tales as animation is put on hold as the cost is exorbitant from Fiji's perspective and the expertise not available locally.

The documentation and storage of data on language, culture and heritage undertaken by government and non-government institutions alike are often faced with challenges as there is a lack of technical expertise and know-how related to the appropriate information and knowledge management technology. The development of systems to be used with databases, servers, user interfaces, back-end to name a few including their maintenance and upkeep are all too new where small island nation language documentation and preservation is concerned. These are pressing issues faced for instance by the Institute of iTaukei Language and Culture, the Fiji Museum, the National Trust of Fiji, Fiji Archives and the Department of Heritage and Arts. The question arises as to whether one digital technology system could feasibly accommodate all the data in one small island nation to reduce labour and costs as after all these institutions collect data from the same indigenous minority groups.

5. Conclusion

Funding is definitely not readily available for MIL to use language technologies. Linguists and language technologists for MIL are hard to come by in small island countries and when they are found, keenness and commitment to the projects by foreign experts do not last. Project owners and funders need to be realistic as well so that projects meet their objectives including the sustainability of projects. The management and funding of technologies in a number of institutions serving the same indigenous communities ought to be considered so that an ideal solution is found and one that will benefit both the targeted communities as well as enhancing institutional technological functions. The challenge also rests with MIL speakers and owners as to how much they want their languages to be documented, developed, accessible, visible, relevant and safeguarded.

6. Bibliographical References

Geraghty, P. (1983). The History of the Fijian Languages. Oceanic Linguistics Special Publication 19. Honolulu: University of Hawai'i Press.

Proceedings, pages 259–262
Paris, UNESCO Headquarters, 5-6 December, 2019.

Building Corpora for Under-Resourced Languages in Indonesia

Totok Suhardijanto, Arawinda Dinakaramani
Department of Linguistics Universitas Indonesia, Computer Science Universitas Indonesia
Depok Indonesia, Depok Indonesia
{totok.suhardijanto, arawinda.dinakaramani}@ui.ac.id

Abstract

Indonesia has the second highest language diversity in the world, just under Papua New Guinea (Simons & Fennig 2019). There are 719 recorded regional languages spoken in Indonesia, 13 of which have become extinct (Lauder 2017). Presently, there are three categories of linguistic condition in Indonesia which consist of the national language, regional language, and foreign language. In accordance with the politics of language, the focus of language development in Indonesia lies in the national language, namely Indonesian or Bahasa Indonesia. Bahasa Indonesia is often cited as one of the great success stories of language policy and planning. However, the very success of Indonesian language threatens the other 699 languages in the archipelago (Cohn & Ravindranath 2014). As a result of the intense politics of national language, language resources focus on Indonesian language—even though its quantity and quality are still far behind other prominent world languages. Thus, the development of regional language resources in Indonesia has yet to become a government priority. Meanwhile, the number of regional languages that fall into the endangered category rises with each passing year.

Keywords: under-resourced languages in Indonesia, multilingual corpora, corpus management system

Résumé

Indonesia merupakan negara kedua yang memiliki keberagaman bahasa tertinggi di dunia setelah Papua Nugini (Simons & Fennig 2019). Tercatat ada 719 bahasa daerah yang dituturkan di Indonesia dan 13 di antaranya telah punah (Lauder 2017). Pada saat ini, di Indonesia, terdapat kondisi kebahasaan dengan tiga kategori bahasa yang hidup di dalam masyarakat, yakni bahasa nasional, bahasa daerah, dan bahasa asing. Sesuai dengan politik kebahasaan, fokus pengembangan bahasa di Indonesia terletak pada bahasa nasional, yaitu bahasa Indonesia. Bahasa Indonesia sering dirujuk sebagai salah satu dari kisah sukses kebijakan pembinaan dan perencanaan bahasa. Namun, kesuksesan bahasa Indonesia tersebut memberikan tekanan terhadap 699 bahasa lain di negara kepulauan tersebut (Cohn & Ravindranath). Akibatnya gencarnya politik bahasa nasional, pengembangan sumber daya bahasa (language resources) pun terfokus pada bahasa Indonesia—meskipun jumlah dan kualitasnya pun masih sangat jauh dari bahasa-bahasa utama di dunia. Dengan demikian, pengembangan sumber daya bahasa daerah di Indonesia belum menjadi prioritas pemerintah, padahal jumlah bahasa daerah yang masuk ke dalam kategori terancam punah terus bertambah tiap tahun.

1. Background

Although Indonesia is the second country with a variety of languages in the world, the documentation effort and development initiatives of language resources from the existing languages in the country are still far from the optimal condition. Even if there are any, according to Suhardijanto & Dinakaramani (2018), all are related to the following two conditions. First, most language resources and documentation were in fact initiated by foreign institutions or institutes. Second, if conducted by Indonesians, they are usually sporadic, individual, and limited. They are limited to be used to support their own research.

Out of 719 languages in Indonesia, 706 of them are still used, while 13 languages belong to the endangered category (Lauder 2016). Regional languages in Indonesia vary in the types of languages and the number of the users. From the types, Indonesian languages can be grouped into two big categories: Austronesian languages and Non-Austronesian languages. The Austronesian languages are spread in the western and eastern parts of Indonesia, while the non-Austronesian ones are spread only in the areas of Papua, Maluku, and Nusa Tenggara, all of which are located in the eastern part of Indonesia. From the number of the users, 386 languages are spoken by more or less 5,000 users; 233 are owned by more or less 1,000 users; 169 are owned by more or less 500 users; and 52 languages belong to more or less 100 users (Gordon 2005). Meanwhile, according to Simons & Fennig (2018), there are only 20 languages spoken by more than one million people, including Javanese with the number of users around 84.3 million.

In the case of Indonesia, besides the high number of regional languages, there are some language problems that complicate the situation. In Indonesia, there are three language categories, namely national language, regional language, and foreign language. According to Riza (2008), the development of language resources for languages in Indonesia generally focus on the national language, which is Indonesian language only. This happens due to the lack of attention from the government towards the problems of regional languages in this country (Lauder 2016). As a result, the regional language development and documentation funding is very limited. Several efforts to establish language resources have been done sporadically and without coordination by the researchers having concern for the fate of the regional languages in Indonesia (see Suhardijanto 2017, Suhardijanto & Arawinda 2018).

Not only the government but also the legislative party seems not to prioritize the language problems in Indonesia yet. This is proven from no law draft on regional languages in Indonesia which has been integrated into the national legislation programs (*prolegnas – program legislasi national*) that becomes the duty of the House of Representatives, although the draft has been completed since 2016. According to Lauder (2016), the enactment of the law on the regional languages, in fact, is expected to be able to strengthen the position of the regional languages in Indonesia. It seems that the government and parliament of Indonesia still consider economic and political fields as the main priority of development in Indonesia.

This paper informs the efforts we have made to collect, compile, and build regional language resources in Indonesia with the funding obtained from various sources. In this study, the discussion focus is limited on the development of the corpus management system that becomes one of the phases in the regional language resources development in Indonesia. The corpus management system we are developing has several functions as follows:

1) save the corpus text data of regional languages in a digital form;
2) process and store corpus metadata so that it can be accessed by other software;
3) analyze corpus text data by corpus methods, such as keyword lists, concordances, n-grams, etc.

2. Corpus Manager

In the effort to build language resources, there are many activities that we have done. Those activities start from language documentation in the field to annotated corpus compilation that can be used to develop the next application of NLP (Natural Language Processing). We started from database development for making dictionaries and grammar books to the making of software or tools to manage language data.

As previously mentioned, this paper will discuss the development of the corpus management system for the existing languages in Indonesia. Since the database or corpora of regional languages are kept in the server of Universitas Indonesia, the corpus system developed is named Korpus Universitas Indonesia (Corpus of Universitas Indonesia).

Corpus management system (CMS) is generally a search engine system developed in a complex manner so that it can carry out the search towards the form of a language or a set of sentences. In a narrow understanding, CMS refers to the server or corpus query engine, while the client side is usually called as user interface (Kouklakis 2007). In this paper, CMS is understood as the combination between those two sides. Therefore, with that understanding, CMS hereinafter will be referred to as a corpus manager.

A corpus manager can be a stand-alone software installed on the user's computer or an online corpus tool that allows users to access corpus, or corpora, from any computer. A corpus manager is designed to have some features. The basic feature is concordance. A user can use a corpus tool to search for a keyword and then the search results will be shown as the line of context for each occurrence of the keyword. Other features of a corpus tool include the ability to extract wordlist, lexical bundles or n-gram, keywords, particular structures, and also metadata information from the corpus.

Some corpus managers are designed for a particular corpus, while other corpus managers are designed to enable users to upload and analyze any corpus. Most corpus managers are used to access a prepared corpus, while some corpus managers are used to access a web as a corpus. Prepared corpus is a corpus that has been compiled with linguistic research in mind and specifically designed for linguists' purposes (Kilgariff & Kosem 2012). The web can be viewed as a corpus with vast quantities of texts for many languages that covers a wide range of text types and domains (Kilgariff, Baisa, Buta, Jakubichek 2014).

The users of corpus managers can be categorized into several types, such as lexicographers, linguistics researchers and students, and language teachers and learners. For this reason, a corpus manager should be designed to meet the needs of their target users. In the case of corpus manager development, one of the designs is the provision of as many as functionalities as possible to fulfill the users' needs. In relation to language resources, the users' needs of this corpus manager are not limited to the researchers in linguistics or other social sciences but also to the researchers and development in the field of natural language processing or artificial intelligence.

3. Language Corpora

Currently, the regional language data that become the focus in the language resources development are limited to the regional language data with the number of users above five million people. The languages entered into Korpus Universitas Indonesia cover Indonesian language, Javanese, Sundanese, Minangkabau language, Banjar language, and Bataknese.

The design of corpus data for our web-based application is decided by considering:

1) selection criteria: if applicable, we design a corpus that represents various texts or genres;
2) corpus size: the size is still growing;
3) data authenticity: from real data, no artificial data
4) storage media: each corpus data is digitalized, especially in the form of text file;
5) data manipulation: we build a web-based application to access and manage corpus data.

Among the six languages, the Javanese corpus possesses the most diverse text collection. Broadly speaking, it is divided into two categories, namely the spoken and written corpus, while the details can be seen in the following figure

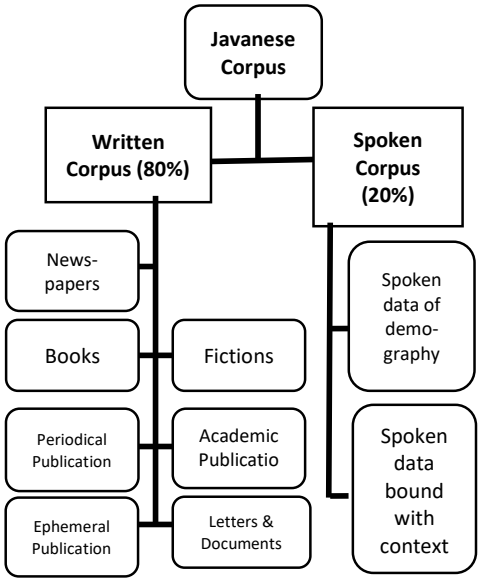


Figure 1: The Structure of Javanese Corpus

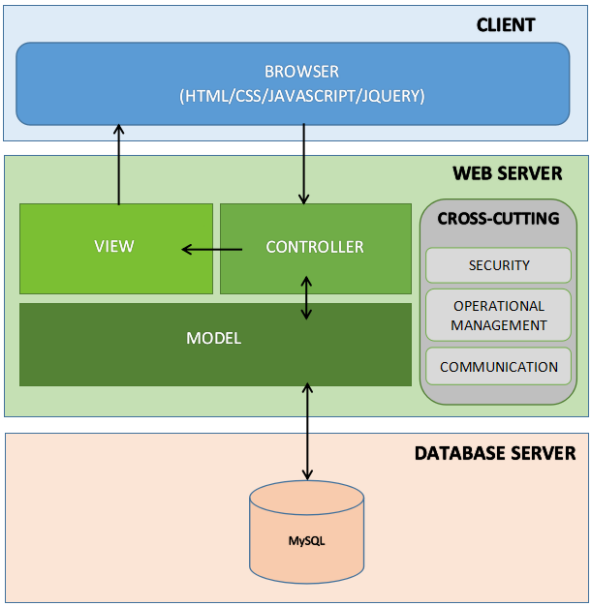


Figure 2: The System Architecture

In the phase of data collection, written texts are generally obtained through data exploration in the field. Data in the form of hardcopy are digitalized via the scanning process and kept in the format of a text file (.txt) with the text encoding UTF-8. Some texts were obtained through text scrapping in the web, such as the Wikipedia text in Javanese.

Meanwhile, for spoken data, they are generally obtained via live recording. Therefore, the process of transcription, editing, and conversion into the format of text file are required as written text data should be. In terms of content, spoken data consist of spoken language data uttered based on demographic variables, such as age, sex, and others. Moreover, other data are spoken language data compiled based on the variety of context, such as lecture, preach, conversation, and others.

Still on the Javanese language corpus, the text data were collected from the period of time between 1940 and 2018. The biggest data portion is the fiction texts. It happened because the most publication encountered in the market in that period of time was fiction texts. In terms of Javanese language data, the least number of texts is academic texts due to the policy of using the national language, which is Indonesian language, as the language of instruction in education.

4. Design and Architecture

The corpus manager is built in PHP and designed based on the model-view-controller (MVC) architectural pattern. Since the users are mostly Indonesians, Indonesian is set as the default language for the application interface page.

Because the purpose is to support the regional language documentation, this corpus manager is also designed to be able to facilitate collaborative work in building and managing language resources. In addition, as the analysis tool, the corpus manager must be able to fulfill the users' needs which vary to search and explore language resources.

Since the functions of this corpus manager vary, the user classification consists of eight types, namely admin, chief editor, editor, data contributor, pending member, pending user, uncategorized, and annotator. The categories of data contributor and annotator are accommodated in this system because the work to build language resources really needs the role of those two user categories.

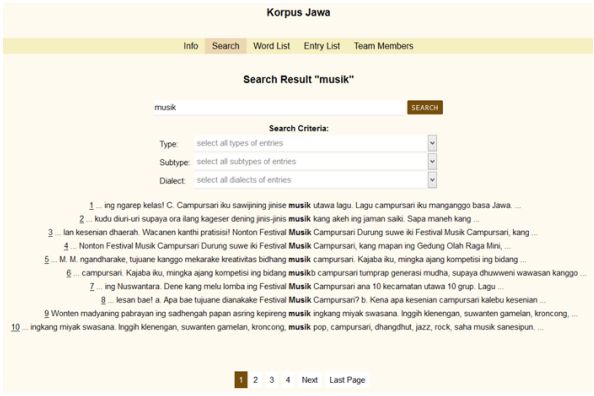
5. Features and Functions

The features in the corpus application have different accessibility permissions. These features are categorized into twofold: accessible to all users and accessible to specific users. Some features that are accessible to all types of users are select corpus, view word list, view concordance search result, and view text list.

Some features that are accessible to a specific type of users are as follows. First, the add editor feature can only be accessed by a chief editor. The submit entry feature can only be accessed by corpus team members. Then, the view, edit, validate, and reject entry features can only be accessed by the editor team. The function of viewing and generating multiple word expression or n-gram can only be accessed by registered users. The Add Annotator feature can only be accessed by a chief editor and an editor member. The function to view, edit, validate, and reject annotation results can only be accessed by chief editor and editor member. Finally, the features of processing and managing text-

annotation can only be accessed by the annotator team, chief editor, and editor member.

Figure 3: The Screenshot of Javanese Concordance Feature



From all functionalities expected to exist in this corpus manager, there are only the functions of concordance, generating word-list, and managing corpus covering uploading, editing, and text sending validation from the contributor.

6. Conclusion

This corpus manager will continue being developed in terms of its design, functionalities, and the number of language data managed. There are several functionalities that are not available yet in this system, such as generating n-gram, collocation, searching based on structure, and others. Some of the corpus managers have been available as a stand-alone software, but they have not been integrated into the system. The number of languages integrated into the system will be attempted to keep on increasing.

7. Acknowledgements

This research was supported by The Directorate of Higher Education, The Ministry of Education Republic of Indonesia with Research Grant No. 1/E1/KP.PTNBH/2019 and No: 516/UN2.R3.1/HKP.05.00/2018.

8. Bibliographical References

Cohn, A. and Ravindranath, M. (2014). Local Languages in Indonesia: Language Maintenance or Language Shift? Linguistik Indonesia 32.2: 131-148.

Dinakaramani, A. and Suhardijanto, T. (2019). Building a web-based application for language resources in Indonesia. Journal of Physics: Conference Series 1192 (2): 12-22.

Gordon, Raymond G., Jr. (ed.), 2005. Ethnologue: Languages of the World, Fifteenth edition. Dallas, Tex.: SIL Internasional.

Kilgarriff, A. and Kosem, I. (2012). Electronic Lexicography. Ed. S. Granger and M. Paquot. Oxford: Oxford University Press, pp 83-106.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. and Suchomel, V. (2014). The Sketch Engine: ten years on. Lexicography, 1: 7-36, 2014.

Lauder, M.R.M.T. 2016. Preventing the Extinction of the Regional Languages through Policy Formation. The paper presented at the 9th National Seminar of Mother Tongue IX, "Prevention Strategy for Indigenous Language Extinction", Denpasar, Bali. 26-27 February 2016.

Riza, H. (2008). Resources Report on Languages of Indonesia. The 6th Workshop on Asian Language Resources.

Simons, G.F. and Fennig, C.D. (eds.). (2019). Ethnologue: Languages of the World. 21 st ed. Dallas: SIL International.

Suhardijanto, T. (2016). Developing language resources for under-resourced languages in Indonesia. The paper presented in the International Conference on Knowledge Creation and Intelligence Computing 2016, State Polytechnics Institute of Manado, Manado, Indonesia, 15—17 November 2016.

Suhardijanto, T. and Dinakaramani, A. (2018). Developing Language Resources for Indigenous Languages in Indonesia: Annotated Javanese Corpus Building. Proceeding of Asia Pacific Corpus Linguistics Conference 2018.

Suhardijanto, T. and Dinakaramani, A. (2019). Korpus Beranotasi: Ke Arah Pengembangan Korpus Bahasa-Bahasa di Indonesia. Prosiding Kongres Bahasa Indonesia. Prosiding Kongres Bahasa Indonesia, pp. 339-355.

Kouklakis, G., Mikros, G., Markopoulos, G., and Koutsis, I. (2007). Corpus Manager A Tool for Multilingual Corpus Analysis. Proceedings from Corpus Linguistics Conference. University of Athens: 1–12.

9. Language Resource References

Suhardijanto, T., Puspitorini, D., and Dinakaramani, A. (2019). Javanese Language Corpus. URL: < https://korpus.ui.ac.id/c/korpus_jawa>.

Suhardijanto, T., Puspitorini, D., and Dinakaramani, A. (2019). Minangkabau Language Corpus. URL: < https://korpus.ui.ac.id/c/korpus_minang>.

Suhardijanto, T., Puspitorini, D., and Dinakaramani, A. (2019). Sundanese Language Corpus. URL: < https://korpus.ui.ac.id/c/korpus_sunda>.

Proceedings, pages 263–266
Paris, UNESCO Headquarters, 5-6 December, 2019.

Language is the Carrier of Our Culture : language documentation as revitalisation in Badimaya and Warriyangga

Rosie Sitorus, Jacqui Cook, Peter Salmon
Bundiyarra – Irra Wangga Language Centre
Geraldton, Western Australia
coordinator@irrawangga.org.au, linguist@irrawangga.org.au, via coordinator@irrawangga.org.au

Abstract

Language documentation in Australia started when European colonisers began interacting with Aboriginal and Torres Strait Islander people. The purpose of documentation has changed throughout time, now focusing on preventing further language loss and strengthening language use in communities. Bundiyarra – Irra Wangga Language Centre (BIW) works with seven languages of the Midwest, Murchison and Gascoyne regions of Western Australia. BIW's work is driven by its community, meaning that solely documentation projects are uncommon and projects combining documentation and revitalisation take precedence. We will explore documentation as revitalisation through case studies of projects in two languages under BIW's purview: Badimaya and Warriyangga.

Keywords: Indigenous languages, language documentation, innovative revitalisation techniques

Wanggayi

Ngunhaburra, walybala nhugurrarrija ganyarawu wagabardi. Guwardi, wirribuga waginha yurnu bagalyaninuru waginhiyawu. Language centre nhugaarrarringu ganyarawu wagabardi ngurra bagalya. Language centre gurlgayinhu nganhurranha ganyarajarri waginhiya warrinthi gurlgabaabaarringu. Nganhurra waginhiya thanala thanangu yirra, warri watharnu nganhala. Nganhurra waginhiya nhurragarrala nganhurrawu yirrawu, nhurragarrawu yirrawu bagalyaninuru Badimayawu, nganhurra yirrawu bagalyaninuru Warriyanggawu Thiinmawu.

1. Introduction : language documentation and revitalisation in Australia

Language documentation[1] in Australia started almost as soon as European colonisers arrived and began interacting with Aboriginal and Torres Strait Islander[2] people. For many early European arrivals, this documentation process was necessary to communicate with the original inhabitants of the continent. L anguage documentation as both an academic sub-field as well as practice has developed to suit the needs of the time, place and people, and since the 1960s and 1970s, it has developed in tandem with language revitalisation efforts. Now, through the advent of language centres and a concerted effort of communities to reclaim their languages, the description and production of materials in endangered languages (as most Australian languages are designated (HRSCATSIA 2012:1)) has become a part of the language revitalisation process (Shulist and Rice, 2019:36).

This paper seeks to examine the changing context in which language documentation has taken place previously, and the emerging context in which it takes place in the present day. We will use two case studies, both projects of the Bundiyarra – Irra Wangga Language Centre, to demonstrate the evolution of language documentation as revitalisation, and to demonstrate how inquiry-driven academia and language community members can work together to conduct research and create resources that would not be possible without such collaboration.

1.1 Context of language documentation in Australia

There has been discussion about the commodification of languages in the context of language documentation (Dobrin, Austin and Nathan, 2009; Shulist and Rice, 2019, amongst others), which centres around the treatment of language as something to be exploited, divorced of the context in which it is given by speakers and commodified for the purposes of "grant-seeking and standard-setting" (Dobrin, Austin and Nathan, 2009:41). The ideologies that yield this commodification can be observed in the way language documentation projects are often assessed, where quantifiable linguistic information such as number of native speakers or levels of fluency supposedly indicate the success of a revitalisation project that does not purport to improve either of these things (Shulist and Rice, 2019).

The influence of the socio-political context in which the need for language documentation has arisen in Australia cannot be understated, as this underpins the dynamic of the relationships between language speakers and documenters. Indigenous Australians face significant barriers to equal participation in Australian society: an average 10.7 year gap in the life expectancy between Indigenous and non-Indigenous Australians demonstrating one, very serious barrier (The Lowitja Institute, 2019). Ongoing disadvantage in health, education and other outcomes presents a demonstrable inequality in power

[1] This paper uses the term 'language documentation' generally to mean the process of language collection and analysis; it is not solely the academic field as discussed by Austin (2013), Shulist and Rice (2019) and others, though that meaning may be included in the discussion in this paper.

[2] The authors have elected to use the terms 'Aboriginal' to denote the Indigenous population of mainland Australia and 'Indigenous' to denote Aboriginal and Torres Strait Islander people collectively throughout this paper.

between Indigenous communities and the often non-Indigenous people working to document their languages.

While much of Australia's language policy nowadays is largely *de facto* and ostensibly aimed at supporting the rights of Indigenous people, historically, policies affecting Indigenous people and the use of Indigenous languages in Australia were outright assimilatory (Truscott and Malcolm, 2010). Years of direct and indirect policies discouraging or banning the use of these languages continue to wield an effect on remaining speakers, many of whom are reluctant to speak their languages outside of familial settings, including with documenters, for fear of punishment or ridicule. For many language communities, including our own, it is a struggle even to document languages, let alone revitalise them.

1.2 Language documentation *and* revitalisation

The notion that language documentation can itself be an act of revitalisation is not a new one (see, for example, Shulist and Rice, 2019). The production of a technical document, such as a grammar, or a dictionary, does not in itself constitute revitalisation, and there remains a gap between documentation and reviving the language in a community of speakers (Shulist and Rice, 2019). Revitalisation requires a different skill set from documentation because it is "about people, not language" (Gerdts, 2017) and so a different approach is needed. Models for integrating the strengths of documentation for the purpose of revitalisation have "emphasized "collaboration" and "participation"…to bridge the gap between the technical goals of academic linguists and the practical needs of speakers (or would-be speakers) of endangered languages" (Shulist and Rice, 2019, discussing K. Rice (2013) and Czaykowska-Higgins (2009)). In Australia, many Indigenous language centres (and Indigenous-led language programs[3]) are leading the integration of documentation and revitalisation.

1.3 Indigenous language centres

Australian Indigenous language centres emerged as the political climate in the late 1980s and early 1990s changed to recognise Indigenous peoples' rights to use their languages (McConvell and Thieberger, 2001:31). Curiously for the authors, language centres as we understand them are an Australian innovation (McConvell and Thieberger, 2001:3; Truscott 2014), borne from the confluence of social empowerment of Indigenous communities, political recognition of the rights of Indigenous people and the timely availability of government funding.

All language centres are different, being by their nature responsive and responsible to their communities. Some have an extensive history of documentation and production of technical resources such as grammars; others are focused on community-driven revival or education as a vehicle for language transmission (Truscott, 2014: 386).

However, despite differences in approach, most language centres share some common experiences. Most are a conduit between their communities and the complex web of policies that provide funding for language projects. Most are subject to the ups and downs of community life, including losing elderly language speakers, dealing with unrest or disagreements in the wider community and working with communities struggling with disadvantage and trauma. Perhaps most importantly, most language centres document and store not only languages, but also "rare and valuable historical materials to which family members may have access" (Truscott 2014: 385). For many communities, language centres have photos, audio recordings and even videos of their Elders[4] and deceased relatives; highly precious material for many communities who face continual loss of language and knowledge as old people, who through much of their lives were forbidden by Australian government policy to pass their knowledge onto their children and grandchildren, pass away. Language centres, then, become repositories of the history and knowledge of communities – a heavy burden, and often one which goes unacknowledged in the broader context of language work.

2. Bundiyarra – Irra Wangga Language Centre

The Bundiyarra Irra Wangga Language Centre (BIW) is a regional language centre located in Geraldton, Western Australia. BIW services the Midwest, Murchison and Gascoyne regions of the State, an area of approximately 500,000 square kilometres (Truscott, 2014). It is funded by the Department of Communication and Arts Indigenous Languages and Arts program to support seven languages – Badimaya, Maglana, Ngarlawangga, Nhanda, Wajarri, Warriyangga and Yinggarda – and has at times supported language work and research on other languages, where requested and/or funded by the community. Each of these languages are critically endangered, and whilst a survey of language speakers in the region has not been conducted for some time, anecdotal evidence suggests at least three of these languages no longer have any living full speakers, with the remaining languages having between one and 50 speakers remaining.

The BIW linguist is the primary documenter at the centre, meaning they are often the ones with whom language speakers spend the majority of their time and effort, and who, in this setting, often conceptualise and deliver language projects (see a great summary of this experience in Truscott, 2014:403-404).

Language speakers are often elderly, with competing demands on their time from family, work,

[3] Language centre vs language program is used to distinguish dedicated organisations whose programs are primarily language documentation and revitalisation activities from language activities that are programs of more general organisations, such as Registered Native Title Bodies Corporate, schools or cultural centres.

[4] 'Elder' is a term with specific meaning for Aboriginal and Torres Strait Islander people. While it often refers to someone who is *elderly*, it more specifically refers to someone with accepted knowledge and experience of cultural matters who can provide guidance to other people within the community.

deteriorating health and cultural obligations, amongst other things. As such, their interactions with the language centre are often limited in some way, meaning that each interaction is critical. The urgency to both record and harness their knowledge has exponentially increased in the last 10-20 years, as speakers have passed away and prized knowledge has passed with them. The relationships language speakers form through documentation work naturally lead to the design of projects that they find interesting and engaging, and with time pressure, these projects become hybrid and wide-ranging.

It is with these pressures in mind that BIW operates to preserve knowledge of those who have and wish to share it, whilst also ensuring that the knowledge does not disappear. It is through this circumspect, relationship-based documentation process that BIW works to revitalise languages, and it is in this setting that the below case studies were developed.

3. Nganang Badimaya Wangga

Nganang Badimaya Wangga[5]: Yarns with Gami Ollie George (NBW) was a project some 20-years in the making. Mr O. George (dec) was a Badimaya man from Kirkalocka Station, near Mount Magnet in the Murchison region of Western Australia. For 20 years, Mr George worked with language centres and linguists to document Badimaya language for (in his own words) "*his children and grandchildren*".

Much of the documentation of Badimaya was done between 2011 and 2014 by Mr George and former BIW linguist James Bednall, with whom Mr George forged a strong relationship. This intensive documentation process, which resulted in the publication of the Badimaya dictionary and topical wordlist book, required Mr George to recall detailed and complex language alone (other Badimaya language speakers, most of whom were Mr George's family, had passed away years earlier). In order to guide Mr George through that process, James would ask questions about his life, and Mr George would in turn tell fascinating stories and anecdotes spanning his childhood, adolescence and working life. Upon completion of the dictionary and wordlist book, it became clear that BIW had amassed a precious collection of stories about Mr George's life, as well as Badimaya culture, local knowledge and national history.

3.1 Designing with community for revitalisation

In discussions about what to do with these stories, it became clear that Mr George hoped to leave a tangible legacy, and showed immense interest in the production of a resource that might transmit his significant linguistic and cultural knowledge to his descendants. Mr George's health had begun to deteriorate (he was 81 when he began working on NBW) and so the BIW project team agreed that this project should capture all facets of his knowledge, from as many angles as possible, to create, as Austin describes, a "multipurpose record…[that] is multi-disciplinary and draws on theoretical concepts and methods" from a variety of fields (2013).

Mr George had previously been the subject of short films about his life and knowledge and had enjoyed the process of filming and seeing his stories in film, a format he observed to be more appealing to his target audience (his descendants). His animated storytelling style also created a rich visual mindscape of long-passed events and people, which were of course impossible to film.

BIW had developed a good working relationship with several local creatives, and so enlisted them to assist in capturing these other dimensions of Mr George's knowledge. Chris Lewis, a filmmaker with the local Australian Broadcasting Corporation station, created a short film about NBW and Mr George, and Brendan Penzer, an experienced curator and visual artist who had worked in the remote township where Mr George lived, coordinated the participation of local artists to create 22 paintings representing some of the stories Mr George told.

At the completion of the project, NBW had recorded approximately 20 additional hours of audio as well as subsequent additional linguistic information, including new entries in the Badimaya dictionary database, created a book depicting the stories told by Mr George in Badimaya and English (printed in sound[6]), produced an art exhibition that has since travelled across the country and created a legacy for Mr George and his family that has continued to restore prestige and pride in Badimaya language.

3.2 NBW to WOC : designing Warriyangga on Country

This approach has become a hallmark of BIW's method of designing projects: observing and understanding the unique abilities, preferences and interests of language speakers, as well as their goals for recording their languages; working with its extended network to coordinate a project team with the specialist skills to fully capture the knowledge of those speakers; and using the materials recorded, create resources for use by the intended audience of the language speaker. Following the success of NBW, BIW has worked to replicate this model of responsive language project design with other language speakers.

In 2017, BIW began working with Warriyangga Elder and language speaker, and co-author of this paper, Peter Salmon. Mr Salmon had not participated in language work before and his interest in sit-down, tedious documentation tasks was limited. However, his knowledge of language was and is unparalleled, except perhaps by his knowledge of country and culture. Well into his 80s, Mr Salmon was interested in documentation tasks that focused on narratives, both autobiographical and cultural, describing the natural features of Warriyangga country, or

[5] Translation: My Badimaya Language

[6] BIW works with PrintingAsia, a printing company that prints books on specially coded paper that is then read by an audio reading device resembling a pen, enabling text to be accompanied by the original audio that has been transcribed.

telling stories of his working life – tasks that, to him, represented a recording of knowledge he wanted to pass onto his descendants. The parallels between the knowledge, the goals and the potential of both Mr George and Mr Salmon were significant and encouraging.

BIW has considered these preferences in co-designing (with Mr Salmon) a documentation and revitalisation project that builds off the successful model exemplified in NBW. The project, Warriyangga on Country (WOC), commenced in October 2019, and brings together a cross-disciplinary team of linguists, an anthropologist, an ethnobotanist, a photographer and videographer to record his knowledge in a variety of media, for the creation of both project materials and a database of the vast knowledge he has of his language, culture and country, which will be available for his descendants long after he has passed away. This team will work with a multimodal team of visual artists and a book designer who will create project materials including an exhibition and a storybook. Perhaps most significantly, WOC will include members of Mr Salmon's family in documentation, helping to demystify the process and building in community ownership of the project and materials produced; a key part of the method to bridge the gap between linguists and would-be speakers (Shulist and Rice, 2019).

4. Conclusion

The best model of ethical language documentation in the community in which BIW works is one that combines documentation with responsive, community-driven revitalisation projects. Many Australian languages face losing their last speakers in the coming decade, and such a model, if used in other settings, would continue documentation and expand revitalisation efforts Documentation can continue, whilst increasing the capacity of the community around speakers whose knowledge is being documented. This model by continuing knowledge transfer between generations, contemplates a future where members of that community become the speakers who are in turn documented themselves. We look forward to it.

5. Acknowledgements

The authors wish to acknowledge the LT4All Organising Committee for their generous support in allowing Ms Sitorus to attend and submit this paper. Ms Sitorus and Ms Cook acknowledge their co-author, Mr Salmon, for his ongoing commitment to preserving his Warriyangga language, culture and country knowledge for future generations, and pay their respects to Mr O. George (dec), who worked so tirelessly on his Badimaya language to provide a rich and well-resourced body of language knowledge for his grandchildren and beyond. Finally, the authors wish to acknowledge the language communities of the Midwest, Murchison and Gascoyne for their unwavering dedication to preserving their languages and ensuring Australia's future linguistic diversity.

6. Bibliographical References

Austin, P. (2013). Language documentation and meta-documentation. In Jones, M. & Ogilvie, S. (Authors), Keeping Languages Alive: Documentation, Pedagogy and Revitalisation, pages 3-15). Cambridge: Cambridge University Press.

Czaykowska-Higgins, E. (2009). Research models, community engagement, and linguistic fieldwork: Reflections on working within Canadian Indigenous communities. *Language Documentation & Conservation*, 3(1):15-50.

Dobrin, L., Austin, P. and Nathan, D. (2009). Dying to be counted: the commodification of endangered languages in documentary linguistics. In Austin, P. (Ed) *Language Documentation and Description*, 6:37-52 London: SOAS.

Gerdts, D. (2017). Indigenous Linguists: Bringing Research Into Language Revitalisation. *International Journal of American Linguistics*, (83(4):607-617.

McConvell P. and Thieberger, N. (2001). State of Indigenous Languages in Australia – 2001. Australia State of the Environment Second Technical Paper Series (Natural and Cultural Heritage). Department of Environment and Heritage, Canberra.

The Lowitja Institute (2019). Close the Gap report - Our Choices Our Voices. Accessed from https://www.humanrights.gov.au/our-work/aboriginal-and-torres-strait-islander-social-justice/publications/close-gap-report-our .

Shulist, S. and Rice, F. (2019). Towards an interdisciplinary bridge between documentation and revitalization: Bringing ethnographic methods into endangered-language projects and programming. *Language Documentation & Conservation*, 13:36-62.

Truscott, A. (2014). When is a linguist not a linguist: the multifarious activities and expectations for a linguist in an Australian language centre. *Language Documentation & Conservation*, 8:384-408.

Truscott, A. and Malcolm, I. (2010). Closing the policy-practice gap: making Indigenous language policy more than empty rhetoric. In Hobson, John and Lowe, Kevin and Poetsch, Susan and Walsh, Michael (eds.), *Re-awakening languages: theory and practice in the revitalisation of Australia's Indigenous languages*, 6-21. Sydney University Press.

House of Representatives Standing Committee on Aboriginal and Torres Strait Islander Affairs (2012), Our Land, Our Languages: Language Learning in Indigenous Communities, Australia.

Proceedings, pages 267–271
Paris, UNESCO Headquarters, 5-6 December, 2019.

Lessons learned after development and use of a data collection app for language documentation (Lig-Aikuma)

Laurent Besacier[1], **Elodie Gauthier**[2], **Sylvie Voisin**[3]
[1]LIG, Grenoble, France [2]LORIA, Nancy, France [3]DDL, Lyon, France

Abstract

Lig-Aikuma is a free Android app running on various mobile phones and tablets. It proposes a range of different speech collection modes (recording, respeaking, translation and elicitation) and offers the possibility to share recordings between users. More than 250 hours of speech in 6 different languages from sub-Saharan Africa (including 3 oral languages in the process of being documented) have already been collected with Lig-Aikuma. This paper presents the lessons learned after 3 years of development and use of Lig-Aikuma. While significant data collections were conducted, this has not been done without difficulties. Some mixed results lead us to stress the importance of design choices, data sharing architecture and user manual. We also discuss other potential uses of the app, discovered during its deployment: data collection for language revitalisation, data collection for speech technology development (ASR) and enrichment of existing corpora through the addition of spoken comments.

1. Introduction

Mobile apps can be now easily produced and authors such as (Drude et al., 2013) believe that an upcoming technological revolution is on the way and that we could face *"a great transformation of the field triggered by an exponential increase in the use of smartphones and tablets .../... even in less developped regions of the world. .../... The development of simple and intuitive app interfaces for smartphones and tablets is having a democratizing effect, allowing for the engagement of user groups who were unable to participate in earlier phases of the digital revolution."*

Using apps on mobile devices, it is now possible to collect audio and video recordings from large number of speakers with lower supervision of a researcher. Apps lower the pressure of defining the best sampling selection process, which speakers and what data exactly to collect. Moreover, additional meta informations can be collected automatically from mobile devices (geographic coordinates, movement patterns, images, time codes). For instance, images (photos taken by potentially hundred of users) can be used to enrich lexical databases or, conversely, these images can be used to elicit speech.

With such a technology, we may envision oral language documentation collections growing very large with many speakers and material to study a bunch of linguistic phenomena, from acoustic-phonetics to discourse analysis, including phonology, morphology and lexicon, grammar, prosody and tonal information. Large scale data collection also allows to collect statistically significant data, for instance on dialectal and socio-linguistic variation. However, large data collections require well organized repositories to access the content, with efficient file naming and metadata conventions that should also facilitate automatic processing.

Contribution. This paper presents the lessons learned after 3 years of development and use of a data collection app (LIG-AIKUMA). While significant data collections were conducted, this has not been done without difficulties. Some mixed results lead us to stress the importance of design choices, data sharing architecture and user manual. We also discuss other potential uses of the app, discovered during its deployment: data collection for language revitalisation, data collection for speech technology development (ASR) and enrichment of existing corpora through the addition of spoken comments.

2. The app and its evolution

LIG-AIKUMA is an improved version of the Android application *Aikuma* initially developed by Steven Bird and colleagues (Bird et al., 2014). Features were added to the app in order to facilitate the collection of parallel speech data in line with the requirements of several field linguists interviewed. The resulting app, called LIG-AIKUMA, runs on various mobile phones and tablets and proposes a range of different speech collection modes (recording, respeaking, translation and elicitation).

The application LIG-AIKUMA has been successfully tested on different devices (including Samsung Galaxy SIII, Google Nexus 6, HTC Desire 820 smartphones and a Galaxy Tab 4 tablet). It can be downloaded from a dedicated website.[1]

Table 1 presents the main features of the app. *Recording* lets simply record speech. *Respeaking*, initially introduced by Woodbury (Woodbury, 2003), involves listening to an original recording and repeating what was heard carefully and slowly. This results in a secondary recording that is much easier to transcribe later on (transcription by a linguist or by a machine). In this recording mode, parallel audio data mapping is captured (between source recording and respoken recording). *Translating* is a translation of an original recording. In *Elicitation* mode, the user can load a text, an image or a video from the device and then record read speech or comment on images/videos.

In addition to those recording modes, the app has the following features: smart generation and handling of speaker metadata (age, languages spoken, geolocalisation) ; automatic backup of interrupted sessions ; data sharing between users ; automatic generation of a consent form (from speaker's metadata) ; export to Elan software.[2] The inter-

[1] `http://lig-aikuma.imag.fr`
[2] `https://tla.mpi.nl/tools/tla-tools/elan/`

Table 1: Main features of LIG-AIKUMA

FEATURES	AIKUMA	LIG-AIKUMA
Recording and documentation	✔	✔
Respeaking and oral translation	✔	✔
Extras : Sync. and Sharing, Geolocalisation, Textless interface	✔	✔
Elicitation (text-image-video) mode	✖	✔
User profiles, Consent form, Metadata	✖	✔
Automatic backup of interrupted sessions	✖	✔
Multilingual interface and User feedback	✖	✔
Documentation (samples, tutorial, ...)	✖	✔
Export to Elan	✖	✔

face and the documentation are available in 3 different languages (English, French and German).

3. Data collection of three oral Bantu languages

So far, LIG-AIKUMA was used to collect data in three unwritten African Bantu languages in close collaboration with three major European language documentation groups (LPP, LLACAN in France; ZAS in Germany).

Basaa, which is spoken by approximately 300,000 speakers (SIL, 2005) from the "Centre" and "Littoral" regions of Cameroon, is the best studied of our three languages. The earliest lexical and grammatical description of Basaa goes back to the beginning of the twentieth century (Rosenhuber, 1908) and the first Basaa-French dictionary was developed over half a century ago (Lemb and de Gastines, 1973). Several dissertations have focused on various aspects of Basaa (Bot ba Njock, 1970; Makasso, 2008) and the language also benefits from recent and ongoing linguistic studies (Dimmendaal, 1988; Hyman, 2003; Hamlaoui and Makasso, 2015).

Myene, a cluster of six mutually intelligible varieties (Adyumba, Enenga, Galwa, Mpongwe, Nkomi and Orungu), is spoken at the coastal areas and around the town of Lambarene in Gabon. The current number of Myene speakers is estimated at 46,000 (Lewis et al., 2013). The language is presently considered as having a "vigorous" status, but the fact that no children were found that could participate in a study on the acquisition of Myene suggests that the language is already endangered. A basic grammatical description of the Orungu variety (Ambouroue, 2007) is available, as well as a few articles on aspects of the phonology, morphology and syntax of Myene ((Van de Velde and Ambouroue, 2011) and references therein).

Our third and last language, **Embosi** (or alternatively Mbochi), originates from the "Cuvette" region of the Republic of Congo and is also spoken in Brazzaville and in the diaspora. The number of Embosi speakers is estimated at 150,000 (Congo National Inst. of Statistics, 2009). A dictionary (Beapami et al., 2000) is available and, just like Basaa and Myene, the language benefits from recent linguistic studies (Amboulou, 1998; Embanga Aborobongui, 2013).

From a linguistic perspective, the three languages display a number of features characteristic of the Bantu family: (i) a complex morphology (both nominal and verbal), (ii) challenging lexical an postlexical phonologies (with processes such as vowel elision and coalescence, which bring additional complexities in the recovery of individual words), and (iii) tones that serve establishing both lexical and grammatical contrasts.

As shown in Table 2, 239h of speech data in 3 languages were collected with LIG-AIKUMA. The corpus is composed of 65h of recorded speech, 83h of respoken speech and 69h of French translations of the respoken utterances. Speech was also elicitated from images or texts (22h).

Table 2: Overview of data collections for 3 oral Bantu languages made with LIG-AIKUMA

Language	Record.	Respeak.	Translat.	Elicitat.
Basaaá	23h	24h	34h	8h
Mboshi	33h	30h	30h	14h
Myene	9h	29h	5h	x
Total	**65h**	**83h**	**69h**	**22h**

More details on these data collections can be found in (Rialland et al., 2018; Hamlaoui et al., 2018) for Mbochi and Basaa respectively. A subset of 5k utterances in Mbochi was also provided to the community for computational language documentation experiments (Godard et al., 2018) and is distributed by ELRA.[3]

4. Lessons learned

4.1. Recording outdoors and indoors

A frequently asked question from linguists about the mobile app concerns the quality of the recordings obtained. Our experience, after recording several hundreds hours of speech, shows that modern smartphones and tablets are equipped with good microphones that provide good quality recordings for further human or automatic analysis. It is important, however, to control the recording environment and following recommandations can be made on this aspect. First, the smartphone (or tablet) must be placed on a table rather than being handled at the time of the recording, in order to avoid unwanted noises due to phone manipulation. Secondly, recording should be done preferably indoors to limit environmental noises (sounds of children or animals, footsteps, discussions, wind, etc). Even indoors, to prevent reverberation and echo, it is preferable not to stand too close to a wall and, if the room is poorly furnished, the walls should be covered with a heavy fabric to muffle the sound. Should the recording occur outdoors, the

[3]Available for free at: `http://catalogue.elra.info/en-us/repository/browse/ELRA-S0396/`

use of a lapel microphone is strongly recommended in order to be as close as possible to the speech source. Finally, another issue faced with the mobile app use outdoors is the reflection of the sun on the tablet/smartphone screen, which sometimes makes it difficult to view videos or images for the elicitation mode (especially with elderly speakers having vision problems).

4.2. Importance of metadata

Filling in metadata is sometimes considered a tedious task by users. This sometimes resulted in incomplete or too quickly filled forms. A consequence is that missing data is found on return from the field trip. For this reason, we developed lately a new feature called *speaker profiles* which saves all speakers' metadata automatically (when metadata of a new speaker is filled in, it is saved - in a so called *profile* - and can be retrieved with an *import* button). Moreover, the possibility of using speaker metadata to automatically generate consent sheets has been proposed and implemented: a *pdf* file is automatically generated according to a template prepared by the user and then filled in with the speaker's information (age, name, etc.). During the use of the app, we also realized the need to add other metadata / informations *at the end of* a recording session (about recording conditions, speaker behavior, etc.) but this feature does not exist yet in our code and is left for future improvements. Finally, the need to annotate non-speech segments, expressed by the users, lead us to introduce this possibility for the *respeaking* and *translation* modes. The time codes of the non speech segments are then stored in a file placed in the same folder as the recording.

4.3. Limited autonomy of mobile devices

The application runs on a mobile terminal with non infinite autonomy (usage time and memory). It is thus essential to plan recording sessions that do not exceed this autonomy and to plan "rest" periods used to recharge the phone's battery and export the recorded data to a laptop or to an external disk. This is all the more important as the mobile phone (or tablet) may be used for other functions such as capturing images and videos. Consequently, when developing the app, we paid particular attention to optimizing the code to preserve the autonomy of the mobile device. In the event of a phone shutdown or crash in the middle of a recording session, we also provide a complete backup of the session history in order to avoid losing data and to be able to return to the exact point of the current session after recharging the phone.

4.4. Global architecture for data collection

An engineer was responsible for gathering data from all the different linguists in 3 bantu languages, backing it up on a single server and checking its integrity. This process must be facilitated if we want to scale up to 100 languages. For instance, the data collected on each deployed mobile device must be regularly deposited (synchronized) on a backend server that ensures data backup and integrity. Such a global architecture for data collection still needs to be designed. Ideally, we would like an architecture that keeps track of all recordings distributed through N autonomous mobile devices, that addresses internet connectivity issues, that verifies data integrity and facilitates automatic quality control. It should also provide less labour-intensive data uploading and compiling.

4.5. Need for documentation and tutorials

In the first few months of the project, misunderstandings about how to use the app led us to write documentation in three languages (French, English, German). We have also added a video tutorial (in French with English subtitles) to these written documents, as well as a quick introduction to the application in the form of a 90mn practical exercice.[4]

5. Future extensions and opportunities

During the use of LIG-AIKUMA, we discovered several extensions and opportunities, not identified at the beginning of the initial project. These are briefly described in this section.

5.1. Data collection for language revitalisation

For Mbochi language, we also recorded, with the app, 1500 pictures illustrating plants, artifacts, animals and everyday activities to be included later on in an Encyclopedia or to be archived as culturally sensitive or to be included in an image book for language teaching. These pictures (see figure 1) were commented by 2, 3 or 4 speakers. Each comment lasted between 20 seconds to 3 minutes. Each image is therefore associated with a recording corresponding to a discussion, between several speakers, about that image.

Figure 1: Example of local pictures used for speech elicitation. Listening to the corresponding recordings, one clearly distinguishes several repetitions of a word corresponding to the main object of the image (left: *ambamba* ; right: *dongodongo*)

5.2. Data collection for speech technology development (ASR)

Originally intended for language documentation and data collection in the field, our app has also been useful for collecting speech for technological development purposes targeting under resourced languages. For instance, 10h of read speech in Fongbe (spoken especially in Benin, Togo, and Nigeria) were collected using the *elicitation* (from text) mode of the app. The first ever ASR system for this language was trained using this initial corpus (Laleye et al., 2016). Similarly, 7h40 of speech in Amharic (Ethiopia) was recorded after translating the Basic Traveler Expression Corpus (BTEC) under a normal working environment

[4] see `https://lig-aikuma.imag.fr/tutorial/` for more details

(Melese et al., 2017). An ASR system was trained to recognize basic travel expressions in Amharic. These two data collections, accelerated by the use of the app, allowed two African doctoral students to quickly record a dataset for their researches in automatic speech recognition (ASR). An ASR system for Wolof (Senegal) was also developed in (Gauthier et al., 2017) and used for analyzing vowel length contrast in different dialectal variants of Wolof.

5.3. Enrichment of existing corpora through the addition of spoken comments

In discussions with several linguists, we realized that many corpora have already been collected to document the world's languages. These data, which exist in different formats (digital or not), are undoubtedly valuable resources that must be safeguarded and enriched. There is a risk that these corpora will disappear with the linguist who recorded them. We think that the *respeaking* mode of the app could allow the linguist to enrich recordings with an additional tier of spoken comments. If speakers of the language are available, it may also be possible to make them repeat part of the dataset under more favorable acoustic conditions (using the *respeaking* mode).

6. Conclusion

This article summarized three years of development and testing of a mobile application for collecting speech in the field: LIG-AIKUMA. A significant amount of speech could be collected to document three Bantu languages. This shows the potential of the app. However, we also presented, in this article, its limitations and the problems encountered during data collection. We also discussed other uses of the app, discovered during its deployment: data collection for language revitalisation or for speech technology development (ASR), enrichment of existing corpora through the addition of spoken comments, etc. LIG-AIKUMA can be downloaded on `https://lig-aikuma.imag.fr` and its source code is also available on `https://gricad-gitlab.univ-grenoble-alpes.fr/besaciel/lig-aikuma`.

7. Bibliographical References

Amboulou, C. (1998). Le Mbochi: langue bantoue du Congo Brazzaville (Zone C, groupe C20). Ph.D. thesis, INALCO, Paris.

Ambouroue, O. (2007). Eléments de description de l'orungu, langue bantu du Gabon (B11b). Ph.D. thesis, Université Libre de Bruxelles.

Beapami, R. P., Chatfield, R., Kouarata, G., and Waldschmidt, A. (2000). Dictionnaire Mbochi - Français. SIL-Congo, Brazzaville.

Bird, S., Hanke, F. R., Adams, O., and Lee, H. (2014). Aikuma: A mobile app for collaborative language documentation. ACL 2014, page 1.

Bot ba Njock, H.-M. (1970). Nexus et nominaux en bàsàa. Ph.D. thesis, Université Paris 3 Sorbonne Nouvelle.

Dimmendaal, G. (1988). Aspects du basaa. Peeters/SELAF. [translated by Luc Bouquiaux].

Drude, S., Birch, B., Broeder, D., Withers, P., and Wittenburg, P. (2013). Crowdsourcing and apps in the field of linguistics: Potentials and challenges of the coming technology. Technical report, The Language Archive, Max Planck Institute for Psycholinguistics.

Embanga Aborobongui, G. M. (2013). Processus segmentaux et tonals en Mbondzi – (variété de la langue embosi C25). Ph.D. thesis, Université Paris 3 Sorbonne Nouvelle.

Gauthier, E., Besacier, L., and Voisin, S. (2017). Machine assisted analysis of vowel length contrasts in wolof. In Proceedings of Interspeech, Stockholm, Sweden, August 2017.

Godard, P., Adda, G., Adda-Decker, M., Benjumea, J., Besacier, L., Cooper-Leavitt, J., Kouarata, G., Lamel, L., Maynard, H., Müller, M., Rialland, A., Stüker, S., Yvon, F., and Boito, M. Z. (2018). A Very Low Resource Language Speech Corpus for Computational Language Documentation Experiments. In Proc. LREC, Miyazaki, Japan.

Hamlaoui, F. and Makasso, E.-M. (2015). Focus marking and the unavailability of inversion structures in the Bantu language Bàsàá. Lingua, 154:35–64.

Hamlaoui, F., Makasso, E., Müller, M., Engelmann, J., Adda, G., Waibel, A., and Stüker, S. (2018). Bulbasaa: A bilingual basaa-french speech corpus for the evaluation of language documentation tools. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018.

Hyman, L. (2003). Basaá (A43). In Derek Nurse et al., editors, The Bantu languages, pages 257–282. Routledge.

Laleye, F. A. A., Besacier, L., Ezin, E. C., and Motamed, C. (2016). First automatic fongbe continuous speech recognition system: Development of acoustic models and language models. In Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, FedCSIS 2016, Gdańsk, Poland, September 11-14, 2016., pages 477–482.

Lemb, P. and de Gastines, F., (1973). Dictionnaire Basaá-Français. Collège Libermann, Douala.

Paul M Lewis, et al., editors. (2013). Ethnologue: Languages of the World. SIL International, Dallas, Texas, seventeeth edition.

Makasso, E.-M. (2008). Intonation et mélismes dans le discours oral spontané en bàsàa. Ph.D. thesis, Université de Provence (Aix-Marseille 1).

Melese, M., Besacier, L., and Meshesha, M. (2017). Amharic-english speech translation in tourism domain. In Proceedings of the Workshop on Speech-Centric Natural Language Processing, SCNLP@EMNLP 2017, Copenhagen, Denmark, September 7, 2017, pages 59–66.

Rialland, A., Adda-Decker, M., Kouarata, G.-N., Adda, G., Besacier, L., Lamel, L., Gauthier, E., Godard, P., and Cooper-Leavitt, J. (2018). Parallel corpora in mboshi (bantu c25, congo-brazzaville). In LREC.

Rosenhuber, S. (1908). Die Basa-Sprache. MSOS, 11:219–306.

Van de Velde, M. and Ambouroue, O. (2011). The

grammar of Orungu proper names. Journal of African Languages and Linguistics, 23:113–141.

Woodbury, A. C., (2003). Defining documentary linguistics, volume 1, pages 35–51. Language Documentation and Description, SOAS.

Proceedings, pages 272–275
Paris, UNESCO Headquarters, 5-6 December, 2019.

Language Technologies for Regional Languages of France: The RESTAURE Project

Delphine Bernhard[1], Myriam Bras[2], Pascale Erhart[1], Anne-Laure Ligozat[3], Marianne Vergez-Couret[4]
[1]LiLPa, Université de Strasbourg, France, [2]CLLE, Université de Toulouse, CNRS, UT2J, France,
[3]LIMSI, CNRS, ENSIIE, Université Paris-Saclay, F-91405 Orsay, France, [4] FoReLLIS, Université de Poitiers, France
[1]{dbernhard,pascale.erhart}@unistra.fr, [2]myriam.bras@univ-tlse2.fr,
[3]anne-laure.ligozat@limsi.fr, [4]marianne.vergez.couret@univ-poitiers.fr

Abstract

The RESTAURE project (2015-2018) aimed at providing digital resources and natural language processing (NLP) tools for three regional languages of France: Alsatian, Occitan and Picard. These languages belong to different language families and are characterized by heterogeneous sociolinguistic situations. In this paper, we focus on the main challenges faced during the project and detail the solutions that we have implemented for the development and distribution of the resources and tools produced. We also present the main lessons learned from the RESTAURE project.

Keywords: Alsatian, Occitan, Picard, language technologies

Résumé

Le projet RESTAURE (2015-2018) visait à fournir des ressources numériques et des outils de traitement automatique des langues (TAL) pour trois langues régionales de France : alsacien, occitan et picard. Ces langues appartiennent à des familles linguistiques différentes et se caractérisent par des situations sociolinguistiques hétérogènes. Dans cet article, nous nous concentrons sur les principaux défis rencontrés au cours du projet et détaillons les solutions que nous avons mises en œuvre pour le développement et la distribution des ressources et outils produits. Nous présentons également les principaux enseignements tirés du projet RESTAURE.

1. Introduction

France has only one official language, French, but many more regional languages are present on the French metropolitan territory (23 according to Leixa et al. (2014), but there is no consensus on this number). In contrast to French, these regional languages are poorly equipped with linguistic resources and NLP tools. In this article, we present the results of the RESTAURE project[1] (2015-2018) aimed at providing digital resources and natural language processing (NLP) tools for three regional languages of France: Alsatian, Occitan and Picard. It brought together researchers from four French research units located in Strasbourg (Université de Strasbourg – LiLPa), Toulouse (Université Toulouse Jean-Jaurès – CLLE-ERSS), Amiens (Université de Picardie Jules Verne – Habiter le monde) and Orsay (LIMSI).

We will first briefly describe the three regional languages of France included in the project (Section 2.). We will then present some challenges to providing language technologies for these languages (Section 3.). We will also discuss the solutions, based on recent recommendations to improve digital language vitality of under-resourced and minority languages (Soria et al., 2013; Ceberio Berger et al., 2018) (Section 4.). Finally, we will present the main lessons learned from the RESTAURE project (Section 5.).

[1]`http://restaure.unistra.fr/`

2. Description of Alsatian, Occitan and Picard

2.1. Alsatian

The Germanic Alsatian dialects are spoken in the North-East of France. The dialectal domain of High-German dialects in France actually stretches farther than the former Alsace region and encompasses part of the Moselle department. Moreover, the dialectal domain can be decomposed in several areas, with Low Alemmanic, High Alemmanic and Central German Franconian dialects being represented. The Alsatian dialects can be traced back to the 6th century and the linguistic changes brought by the Alemanni and the Franks (Huck, 2015). The last decades have however seen a decline in the use of the Alsatian dialects, with French being used as the main language of communication in the region.

The Alsatian dialects have mainly been used orally, with a small literary production since 1816 (mainly theater plays and poetry). Spelling is not standardized, which, in addition to *spatial variation* (both on the phonological and lexical levels) accounts for the very diverse graphical variants found in writing.

2.2. Occitan

Occitan is a romance language spoken in southern France and in Val d'Aran in Spain and in several valleys of Italy. Occitan has several varieties organized in dialects. The most accepted classification suggested by Bec (1996) includes Auvergnat, Gascon, Languedocien, Limousin, Provencal and Vivaro-Alpin. However, each dialect has

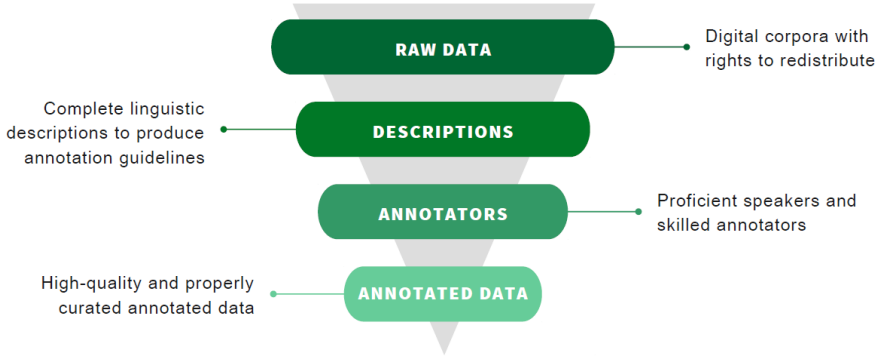


Figure 1: Data bottleneck.

also internal variations. Occitan is written since the Middle Ages and an extensive body of literature has been produced. Although much less socialised than it was before World War II, Occitan is now present in newspapers, on the Internet, on the radio and television, and in some schools and universities.

There are two main spelling standards: the 'mistralienne' spelling designed in the mid-19th by Frederic Mistral and the 'classical' spelling from the 20th century based on medieval conventions whose aim is to minimize the dialectal differences while keeping dialectal particularities (Sibille, 2006). However, literature in Occitan is characterized by a plethora of non-standard individual spellings.

2.3. Picard

Picard is a langue d'oïl (Romance language group) spoken in the North of France (Hauts-de-France) and the Belgian province of Hainaut. Picard has several varieties and spelling is not standardized. Picard is, however, used in writing, as shown by the PICARTEXT database (Eloy et al., 2015), which includes literary works, totalling about 5 million words.

3. Challenges

In this section, we present the most important challenges we have faced during the project.

3.1. Data Bottleneck

Figure 1 sums up what we call the *data bottleneck* challenge for collecting and producing high-quality and properly curated linguistically annotated data. Even if the problems presented are not confined to under-resourced languages, they are even more important for them.

First, collecting raw corpora is made difficult by the scarcity of available resources. For instance, it is usually easy to collect very large corpora on the Web (e.g. using Wikipedia) for languages with many speakers and a good online presence. This is much more of a challenge for under-resourced languages.

Second, accurate and complete linguistic descriptions are needed to enrich corpora with annotations (e.g., part-of-speech, morphosyntactic features). Up-to-date grammars are very difficult to find for regional languages of France: if they exist, they are often outdated or incomplete.

Third, concerning the annotation work *per se*, it is hard to recruit people who are both proficient speakers and skilled annotators.

3.2. Dialectal and Spelling Variation

As already hinted at in Section 2., Alsatian, Occitan and Picard are neither homogeneous nor fully standardized. Different varieties or dialects of these regional languages can be identified in each region. Spelling conventions are either rather recent, or not much used, or even accommodate for dialectal particularities. All in all, dialectal and spelling variation is challenging for NLP tools. For instance, the uncontrolled use of punctuation marks makes it difficult to develop reliable tokenizers, which automatically break down texts into words (Bernhard et al., 2017)

4. Solutions

The solutions we have implemented include a large part of the recommendations by Soria et al. (2013), both for the development of resources and tools (see Section 4.1.) and their distribution (see Section 4.2.).

4.1. Development of Resources and Tools

The development of resources and tools was based on three main principles, in accordance with (Soria et al., 2013): (1) cooperation, (2) use of standards and (3) re-use and recycling of existing tools.

4.1.1. Cooperation

Cooperation between the teams involved in the RESTAURE project was an important asset, all the more so as the different teams specialized in different domains and had variable previous experience in producing language technologies for under-resourced languages. It lead to the development of a common corpus annotation workflow (see Figure 2) and to collaboration in carrying out the various sub-tasks.

For instance, Strasbourg and Toulouse cooperated to perform Optical Character Recognition (OCR) for corpus acquisition (Vergez-Couret et al., 2015). Strasbourg and Amiens worked together to develop a tokenizer for Picard (Bernhard et al., 2017). Orsay provided help to Amiens to parse and format lexicons. Finally, Orsay and Strasbourg

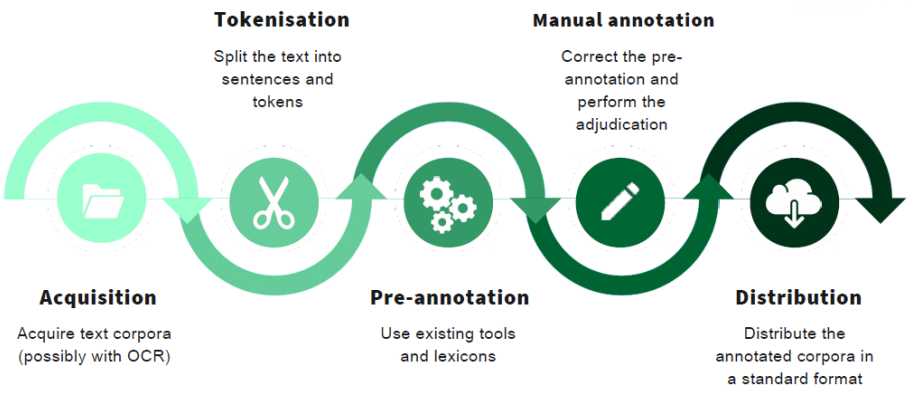


Figure 2: Corpus annotation workflow. Icons made by Tomas Knop, Smashicons, Freepik from www.flaticon.com

cooperated on the task of identifying place names for Alsatian (Bernhard et al., 2018b). Clearly and in retrospect, many tasks could not have been accomplished, or in a less sophisticated form, without the collaboration between the different teams. This cooperation made it possible to compensate, to some extent, for the lack of human resources and specialists for the regional languages under study.

4.1.2. Use of Standards

As Soria et al. (2013) write:

"Use of standards is the key to interoperability of resources, as they allow resource sharing, re-usability, maintainability and long-term preservation."

We thus chose to share the annotated corpora produced by the RESTAURE project in the CONLL-U format, defined in the *Universal Dependencies* (UD) project (Nivre et al., 2016). This format is directly usable for training POS (Part-Of-Speech) tagging tools such as spaCy [2] or UDPipe (Straka and Straková, 2017). Moreover, the Universal POS tags defined in UD helped us define tagsets for Alsatian and Picard as well as write annotation guidelines based on the UD recommendations. The original tagsets for Alsatian, Occitan and Picard are not strictly identical to the UD POS tags, but could be transformed into these tags using a correspondence table. The procedure for transforming our corpora into UD format is described in (Miletic et al., 2019).

4.1.3. Re-Use and Recycling of Existing Tools

During the course of the project, we re-used and recycled existing tools, whenever possible:

- OCR (Vergez-Couret et al., 2015):
  - Tesseract (Smith, 2007)
  - Jochre (Urieli and Vergez-Couret, 2013)
- Part-of-speech (POS) tagging (Vergez-Couret and Urieli, 2015; Bernhard et al., 2018a):
  - for Occitan, Talismane (Urieli, 2013) and APERTIUM (Armentano I Oller, 2008)
  - for Alsatian, TreeTager for German (Schmid, 1994)
- Corpus annotation (Bernhard et al., 2018a): Analog tool (Lay and Pincemin, 2010)

4.2. Distribution of Resources and Tools

The distribution of resources and tools produced during the RESTAURE project also followed three main principles, again in accordance with (Soria et al., 2013): (1) document resources and technologies, (2) be open and (3) share and sustain. The outputs of the RESTAURE project are shared on the Zenodo platform (`https://zenodo.org/communities/restaure`), under a Creative Commons Attribution Share Alike 4.0 licence (CC-BY-SA). The resources and tools are associated with a DOI and are fully documented.

5. Lessons Learned from the RESTAURE Project

Finally, we detail the lessons learned during the course of the project:

Cooperation is key This work on regional languages of France could not have been carried out without real cooperation between various teams with complementary skills (sociolinguistics, dialectology, natural language processing). The parallel work on several languages made it possible to benefit from the experiences carried out on other languages and thus gain in efficiency. The problems that arose in one language led to increased vigilance on this subject in the other languages.

Do not feel inferior to "big" languages Working on under-resourced languages often means starting building language technologies from (or almost from) scratch. It is easy to feel that you are far behind in comparison to better-resourced languages with many more researchers, resources and tools. Producing language resources requires time and the means to do so, and both are rare for under-resourced languages. These extrinsic constraints are difficult to control but should not undermine the desire of researchers to keep working on these languages. This requires that funding agencies as well as program and reviewing committees acknowledge the specific challenges of work on under-resourced languages.

[2] `https://spacy.io/`

Do not reinvent the wheel This is an important principle. First, this means that instead of developing new tools, it is often less time-consuming to try and find a similar tool which can be adapted to your own needs. It is also necessary to learn from similar projects, including e.g. existing guidelines for annotating corpora. Within the project, participants should share a common workflow and use the same tools, if possible. In return, it is important to distribute the resources that have been created, so that the work is beneficial to others.

Focus on data rather than tools Nowadays, most NLP tools are able to learn from data. Methods have evolved from being predominantly based on rules towards machine learning techniques, which are in principle applicable to a wide variety of languages. The main condition is that data are available for re-training them. It is therefore advisable to concentrate on data collection and annotation, rather than on the development of tools. As stressed earlier, tools can then be re-used or re-trained.

6. Acknowledgements

This work was supported by the French "Agence Nationale de la Recherche" (ANR) through the RESTAURE project (no.: ANR-14-CE24-0003).

7. Bibliographical References

Armentano I Oller, C. (2008). Traduction automatique occitan-catalan et occitan-espagnol: difficultés affrontées et résultats atteints. In *IXème Congrès International de l'Association Internationale d'Etudes Occitanes*, Aachen.

Bec, P. (1996). *La langue occitane*. Paris, PUF.

Bernhard, D., Todirascu, A., Martin, F., Erhart, P., Steiblé, L., Huck, D., and Rey, C. (2017). Problèmes de tokénisation pour deux langues régionales de France, l'alsacien et le picard. In *Actes de l'atelier "Diversité Linguistique et TAL" – DiLiTAL 2017*, pages 14–23.

Bernhard, D., Ligozat, A.-L., Martin, F., Bras, M., Magistry, P., Vergez-Couret, M., Steiblé, L., Erhart, P., Hathout, N., Huck, D., Rey, C., Reynés, P., Rosset, S., Sibille, J., and Lavergne, T. (2018a). Corpora with Part-of-Speech Annotations for Three Regional Languages of France: Alsatian, Occitan and Picard. In *Proceedings of the 11th edition of the Language Resources and Evaluation Conference*, Miyazaki, Japan, May.

Bernhard, D., Magistry, P., Ligozat, A.-L., and Rosset, S. (2018b). Resources and Methods for the Automatic Recognition of Place Names in Alsatian. In Andrew U. Frank, et al., editors, *Corpus-Based Research in the Humanities*, volume 1 of *Proceedings of the Second Workshop on Corpus-Based Research in the Humanities CRH-2*, pages 35–44, Vienna, Austria.

Ceberio Berger, K., Gurrutxaga Heraiz, A., Baroni, P., Davyth, H., Kruse, E., Quochi, V., Russo, I., Salonen, T., Sarhimaa, A., and Soria, C. (2018). Digital Language Survival Kit. The DLDP Recommendations to Improve Digital Vitality. Technical report.

Eloy, J.-M., Martin, F., and Rey, C. (2015). PICARTEXT: Une ressource informatisée pour la langue picarde. In *Actes de TALaRE 2015 - Traitement Automatique des Langues Régionales de France et d'Europe*.

Huck, D. (2015). *Une histoire des langues de l'Alsace*. La Nuée bleue, Strasbourg. 24 cm. Bibliogr. p. 447-457.

Lay, M.-H. and Pincemin, B. (2010). Pour une exploration humaniste des textes: AnaLog. In *Proceedings of 10th International Conference Journées d'Analyse statistique des Données Textuelles*.

Leixa, J., Mapelli, V., and Choukri, K. (2014). Inventaire des ressources linguistiques des langues de France. Technical Report ELDA-DGLFLF-2013A.

Miletic, A., Bernhard, D., Bras, M., Ligozat, A.-L., and Vergez-Couret, M. (2019). Transformation d'annotations en parties du discours et lemmes vers le format Universal Dependencies : étude de cas pour l'alsacien et l'occitan. In *Actes de TALN 2019*.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., and others. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49.

Sibille, J. (2006). L'occitan, qu'es aquò. *Langues et Cité : bulletin de l'observation des pratiques linguistiques*, 10.

Smith, R. (2007). An overview of the Tesseract OCR engine. In *Proceedings of the 9th International Conference on Document Analysis and Recognition*, pages 629–633.

Soria, C., Mariani, J., and Zoli, C. (2013). Dwarfs sitting on the giants' shoulders–how LTs for regional and minority languages can benefit from piggybacking major languages. In *Proceedings of XVII FEL Conference*.

Straka, M. and Straková, J. (2017). Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*.

Urieli, A. and Vergez-Couret, M. (2013). Jochre, océrisation par apprentissage automatique : étude comparée sur le yiddish et l'occitan. In *Actes de TALARE 2013 : Traitement Automatique des Langues Régionales de France et d'Europe*, pages 221–234.

Urieli, A. (2013). *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Ph.D. thesis, Université de Toulouse II le Mirail.

Vergez-Couret, M. and Urieli, A. (2015). Analyse morphosyntaxique de l'occitan languedocien : l'amitié entre un petit languedocien et un gros catalan. In *Actes de TALARE 2015 : Traitement Automatique des Langues Régionales de France et d'Europe*.

Vergez-Couret, M., Bernhard, D., Urieli, A., Bras, M., Erhart, P., and Huck, D. (2015). Océrisation de textes pour les langues régionales. Regards croisés sur l'occitan et l'alsacien. In Emmanuelle Chevry Pébayle, editor, *Actes du 10e colloque ISKO France 2015*, pages 250–269.

Proceedings, pages 276–279
Paris, UNESCO Headquarters, 5-6 December, 2019.

Cardamom: Comparative Deep Models for Minority and Historical Languages

John P. McCrae and Theodorus Fransen
Data Science Institute / Insight Centre for Data Analytics
National University of Ireland, Galway
{john.mccrae, theodorus.fransen}@insight-centre.org

Abstract

This paper gives an overview of the Cardamom project, which aims to close the resource gap for minority and under-resourced languages by means of deep-learning-based natural language processing (NLP) and exploiting similarities of closely-related languages. The project further extends this idea to historical languages, which can be considered as closely related to their modern form, and as such aims to provide NLP through both space and time for languages that have been ignored by current approaches.

Keywords: natural language processing, under-resourced languages, deep learning

Achoimre

Tugtar léargas ginearálta sa pháipéar seo ar an tionscadal Cardamom; tionscadal taighde a bhfuil sé mar aidhm aige easnamh acmhainne a laghdú do theangacha mionlaigh agus do theangacha nach bhfuil mórán acmhainní ann ina leith trí phróiseáil teanga nádúrtha, atá bunaithe ar an domhainfhoghlaim, agus trí leas a bhaint as cosúlachtaí teangacha a bhfuil dlúthbhaint acu lena chéile. Áirítear mar chuid den tionscadal seo teangacha stairiúla de bharr go meastar go bhfuil dlúthbhaint acu lena bhfoirm nua-aimseartha. Dá réir sin, tá sé mar aidhm ag Cardamom próiseáil teanga nádúrtha a sholáthar ó thaobh spáis agus ama do theangacha a ndearnadh neamhaird orthu go dtí seo ó thaobh cur chuige an lae inniu de.

1. Introduction

There are estimated to be about 7,000 languages spoken in the world, but currently digital language tools support only a small fraction of these languages. Recent breakthroughs in natural language processing (NLP) have been based on the emergence of deep learning for processing texts and in particular in the use of vectors (*word embeddings*) to represent the meaning of words. Such representations have been shown to be truly interlingual and to allow translation between language pairs without any training data for that pair (Johnson et al., 2017). Deep learning has been enabled by the big data resources for NLP. However, it has been thought to be unsuitable for under-resourced languages as there is insufficient data to train these models. The comparative method, a keystone of modern linguistics (Schleicher, 1876), shows us that careful comparison of closely-related languages can give deep insight into the history, structure and semantics of a language. The key goal of the Cardamom project[1] is the creation of vector-based models of language that take into account the shared phonetic, etymological and semantic information of words in closely-related languages and their application to minority and historical languages. Speakers of minority languages are among the fastest growing communities on the Web and meeting their need is of major societal and commercial importance. Secondly, in order to meet the growing demand for text analysis in digital humanities, whereby access to large corpora text in languages such as Latin, Old English and Old Irish can enable new insights in the study of history and literature, we will develop technologies for historical languages.

The rest of this paper is organized as follows. Section 2. looks at the resource gap that underlies the motivation of the Cardamom project. Section 3. gives a short overview of the state-of-the-art in NLP, constituting the background for the novel methodology employed in our project, discussed in section 4. Advances and opportunities are the subject of section 5., followed by a conclusion in section 6.

2. The resource gap

Providing support for a new language to an existing language technology is by no means an easy goal. Few commercial or public institutes have a clear plan of how to scale beyond 100 languages, as most language technologies can only be developed with experts who speak the language. As such, new methods are required to develop NLP technologies that are viable for real-world applications, in that there is sufficient data and tools to enable enterprises to develop NLP applications for these languages and derive commercial benefits from addressing these minority populations. For example, it is generally considered that over 10 million words (MW) of parallel text are required to train a basic machine translation system, and in OPUS (Tiedemann, 2009)[2] there are only a few languages such as German (919.1MW parallel with English) or Hindi (13.2MW) where this is true. For most languages there is at least a moderate resource gap, such as for Irish (6.9MW), or more frequently a huge resource gap, such as for Scottish Gaelic (0.5MW). Modern algorithms for NLP claim to only require sufficient training data so they can adapt to any task or domain, yet all languages present unique challenges and most real-world systems have required language expertise in their development.

In spite of the large potential impact of the development of language technologies for minority languages, there has

[1] The project website is at `http://cardamom.insight-centre.org/`.

[2] This is based on data collected in 2017.

been comparatively little related focused work on these languages in computer science or linguistics. There are some platforms for the collection of information about under-resourced languages, most notably PanLex (Westphal et al., 2015), Glottolog (Nordhoff, 2012), An Crúbadán (Scannell, 2007) and Ethnologue (Lewis et al., 2009), yet these platforms have not focused on the development of language technologies but instead on language preservation. In contrast, there have been a number of workshops organized in the field of NLP on the topic of 'under-resourced languages', however the focus of these workshops has often been on national languages that have direct support from national funding agencies. For example, a recent report on European languages classified all but 2 EU languages as 'severely' under-resourced (Rehm and Uszkoreit, 2013). Despite the potentially huge impact of work on minority languages, they have been surprisingly neglected and thus there would be significant benefit in providing language technology for these languages. Figure 1 shows a geographic distribution of speakers of digitally emerging languages.

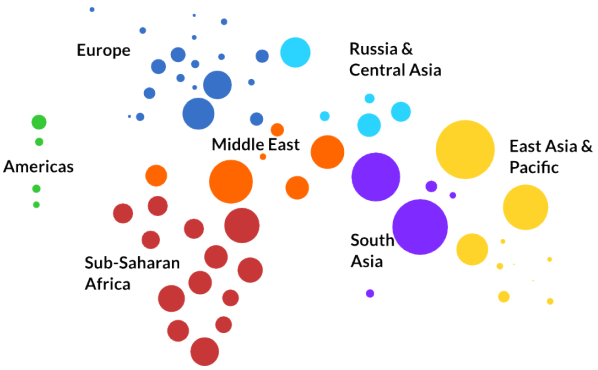


Figure 1: Geographic Distribution of Speakers of Digitally Emerging Languages.

Another important motivation for this project is the growth in digital humanities, where much research is focused on texts written in minority languages, dialects and, most importantly, in historical languages. Distant reading approaches, whereby literary analysis is done by comparing large volumes of text, have given humanities researchers new insights, for example into the linguistic style of a text (*stylometry*). Research in this area draws heavily from NLP techniques and, more recently, deep learning in particular (Brocardo et al., 2017), although a gap still exists between the state-of-the-art in NLP and the tools that are available to digital humanities researchers. Moreover, humanities researchers are often interested in pre-modern texts, where the language and grammar may not correspond to modern languages usage. As such, these texts can be considered to be written in an under-resourced language, which is closely related to an existing modern language. This underlines the interdisciplinary nature of this project and the input and collaboration with researchers in literature and history will be required.

The Cardamom project will address the resource gap with a big data approach to automatically produce resources and technologies for under-resourced languages by exploiting Web content and a deep learning approach grounded in linguistic theory that can leverage a wider range of input data. This will require developing a set of basic NLP approaches, which can build tools for all the world's languages.

3. State-of-the-art in NLP

Deep learning has revolutionized natural language processing technologies and the use of word embeddings such as *word2vec* (Mikolov et al., 2013) has become standard in the field. These methods, along with distributional methods that preceded them, associate a vector with each word, which is also called a word embedding. However, it has been shown that further breaking words up into smaller subword units (Sennrich et al., 2016) or acoustic units (Kamper et al., 2016) can improve quality in tasks such as machine translation. For minority languages, using many languages as a pivot can detect complex morphological phenomena (Asgari and Schütze, 2017). Furthermore, these vectors can very accurately induce semantic similarity (Tai et al., 2015) even in a cross-lingual setting (M^{c}Crae et al., 2013) and this has led to the development of bilingual lexicon induction (Haghighi et al., 2008), where translations are learnt without the need for any existing parallel translations. This work has been extended to full machine translation by a process known as linguistic decipherment (Ravi and Knight, 2011) and such approaches have been shown to enable machine translation systems to be trained on many languages simultaneously (Firat et al., 2017). It is our belief that these state-of-the-art deep models can be employed also for less-resourced languages, both modern and historical. Section 4. will outline the methodology of the Cardamom project in more detail.

4. Methodology

We propose the development of a new model for machine learning over natural languages, that will break the paradigm of learning models independently for each language, but instead learn models for closely-related languages simultaneously. The primary goal of this is to overcome the lack of data for minority and history languages, thus developing new tools and insights for researchers in computer science, linguistics and the humanities. This project will be primarily focused on three key areas: Firstly, as a data science project, we will attempt to find as much information on specific languages in as many forms as possible and combine this using linguistic linked data methods (M^{c}Crae et al., 2013), which have been acknowledged as a key technique for under-resourced languages (Westphal et al., 2015).

Secondly, for deep learning we will apply existing unsupervised methods, such as word embeddings, and develop them further into generic methods that can process minority and historical languages. Thus, we will develop comparative algorithms to exploit similarities between closely-related languages, to overcome the data gap for under-resourced languages. We will approach this first by identifying a small set of about 100 languages in four families of closely-related languages (Celtic, Germanic, Indic and

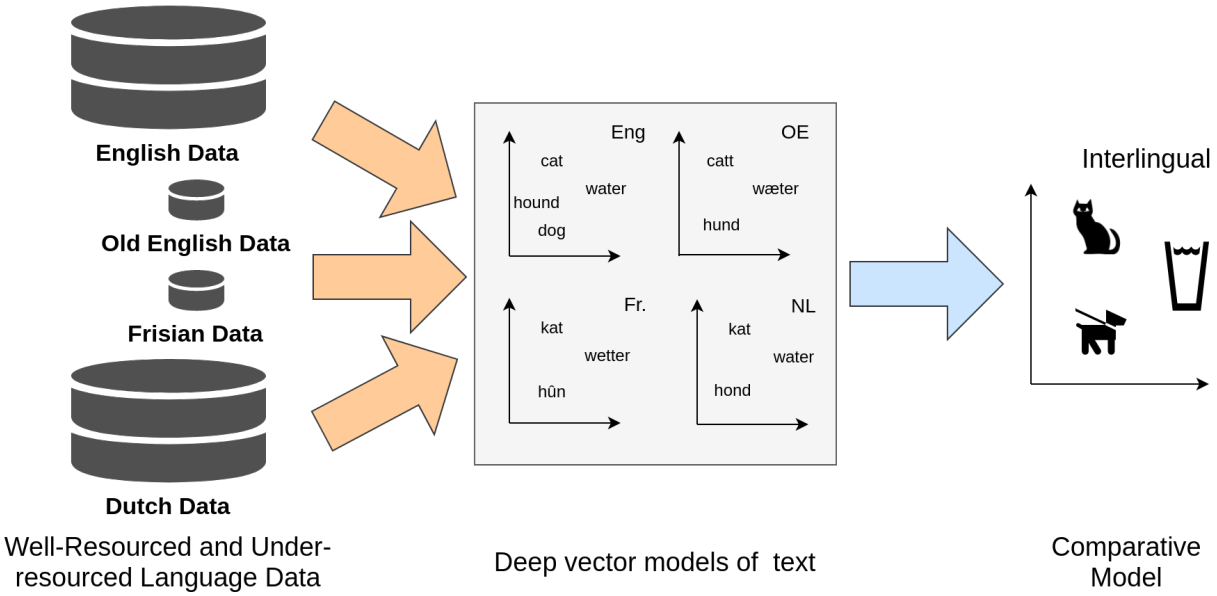


Figure 2: The architecture of Cardamom.

Dravidian) which include a mixture of minority languages (e.g., Irish, Frisian, Tulu) and historical languages (e.g., Old Irish, Old English, Sanskrit). This approach is illustrated in Figure 2, where we see how languages with less data (Old English and Frisian) can be learned simultaneously with well-resourced languages (English and Dutch) to produce a single representation. This model needs to consider orthographic differences ('cat' vs. 'kat'), phonetic changes ('water' vs. 'wetter') and semantic changes (the change of 'hound' in English to refer to only some types of dogs), which will be handled by unsupervised machine learning.

This project aims to revolutionize NLP, which has so far been overwhelmingly applied to major languages such as English, which is a language with comparatively low morphological complexity and standardized spelling and grammar. In contrast, minority languages frequently have complex morphology and significant variation among dialects. As such, the study of such languages raises questions for processing that have not been considered so far. Moreover, the comparative nature of this work requires us not to consider words as independent units, as has typically been done in existing word embedding work, but instead to look into the phonemes that compose the words to establish relationships between dialects (Sennrich et al., 2016).

5. Advances and opportunities

The Cardamom project will open up new research areas in language processing that have not yet been treated. Moreover, the uniquely broad nature of this study covering languages from different areas of the globe as well as different period of times leads to new research opportunities. Given that there are over 2,000 languages being used on the Web and the increasing economic importance of speakers of these languages, it is likely that the unique challenges of these languages are going to become increasingly important for research, societal and commercial applications. This represents a shift in viewpoint that requires the development of new algorithms that tackle problems with novel methods, for example the development of unsupervised morpho-syntactic systems. It is expected that the development of these tasks will provide new viewpoints on existing tasks in NLP. In particular, it is expected that this work will create a shift in approaches in machine translation by providing data in a wide range of languages, spurring development of novel approaches to handle these languages.

The algorithms developed in this project for under-resourced languages will lead to novel developments in the wider context of artificial intelligence in a number of ways: Firstly, the challenges of developing robust algorithms that can work on limited data with a potential highly complex output space (e.g., identifying a wide range of languages) will require the development of novel applications of machine learning. Secondly, the large number of languages studied will make interesting new developments in cognitive sciences by allowing for comparison in, for example, the meaning of words, to be examined in a new and wider situation. Finally, the development of computer-aided language learning software and the primary role of social media will be of interest to researchers in fields such as e-Government, in particular as this work aims to develop interaction with speakers of languages in the G77. Finally, the development of computer-aided language learning technology will have an impact on the field by widening the area of study and providing real case studies on teaching languages that are severely under-resourced. It will also have a wider societal impact in encouraging young learners to adopt languages that are currently mostly only used by older members of their communities.

6. Conclusion

This paper has described the Cardamom project, the aim of which is to use NLP and deep learning applied to a set of minority and historical languages primarily in four language families: Celtic, Germanic, Indic and Dravidian. The methodology involves a big data approach with largely unsupervised models that are simultaneously applied to closely-related languages, in order to overcome the data gap for under-resourced languages. The results are expected to advance the current state-of-the-art computational models and translate into societal and commercial applications.

7. Bibliographical References

Asgari, E. and Schütze, H. (2017). Past, present, future: A computational investigation of the typology of tense in 1000 languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 113–124.

Brocardo, M. L., Traore, I., Woungang, I., and Obaidat, M. S. (2017). Authorship verification using deep belief network systems. *International Journal of Communication Systems*, 30(12):e3259.

Firat, O., Cho, K., Sankaran, B., Vural, F. T. Y., and Bengio, Y. (2017). Multi-way, multilingual neural machine translation. *Computer Speech & Language*, 45:236–252.

Haghighi, A., Liang, P., Berg-Kirkpatrick, T., and Klein, D. (2008). Learning bilingual lexicons from monolingual corpora. In *Proceedings of ACL-08: HLT*, pages 771–779.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., and Dean, J. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Kamper, H., Jansen, A., and Goldwater, S. (2016). Unsupervised word segmentation and lexicon discovery using acoustic word embeddings. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 24(4):669–679.

Lewis, M. P., Simons, G. F., and Fennig, C. D. (2009). Ethnologue: languages of the world, Dallas: SIL International. *Online version: http://www.ethnologue.com*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Workshop Track Proceedings*.

M^{c}Crae, J. P., Cimiano, P., and Klinger, R. (2013). Orthonormal explicit topic analysis for cross-lingual document matching. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1732–1740.

Nordhoff, S. (2012). Linked data for linguistic diversity research: Glottolog/Langdoc and ASJP online. In C. Chiarcos, et al., editors, *Linked Data in Linguistics*, pages 191–200. Springer.

Ravi, S. and Knight, K. (2011). Deciphering foreign language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 12–21.

Rehm, G. and Uszkoreit, H. (2013). *Strategic research agenda for multilingual Europe 2020, presented by the META Technology Council*. Springer.

Scannell, K. P. (2007). The Crúbadán project: Corpus building for under-resourced languages. In *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*, volume 4, pages 5–15.

Schleicher, A. (1876). *Compendium der vergleichenden Grammatik der indogermanischen Sprachen*. Hermann Böhlau, Weimar.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Tai, K. S., Socher, R., and Manning, C. D. (2015). Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1556–1566.

Tiedemann, J. (2009). News from OPUS – a collection of multilingual parallel corpora with tools and interfaces. In N. Nicolov, et al., editors, *Recent advances in natural language processing*, volume 5, pages 237–248. John Benjamins.

Westphal, P., Stadler, C., and Pool, J. (2015). Countering language attrition with PanLex and the Web of Data. *Semantic Web*, 6(4):347–353.

Proceedings, pages 280–283
Paris, UNESCO Headquarters, 5-6 December, 2019.

Empowering Indigenous Communities through Citizen Linguistics, Language Resources and Human Language Technologies

Christopher Cieri, Mark Liberman
University of Pennsylvania, Linguistic Data Consortium
3600 Market Street, Philadelphia, PA 19104 USA
{ccieri, myl|} @ldc.upenn.edu

Abstract

This paper demonstrates the link between UNESCO goals, language technologies and the requirements of language resources. It describes the causes of the scarcity of language resources and proposes a novel method to increase the supply of linguistic data by empowering indigenous language communities to contribute directly.

Keywords: citizen linguistics, language resources, corpora, human language technologies

Résumé

Questo documento dimostra il legame tra gli obiettivi dell'UNESCO, le tecnologie linguistiche e i requisiti delle risorse linguistiche. Descrive le cause della scarsità delle risorse linguistiche e propone un nuovo metodo per aumentare la fornitura di dati linguistici dando potere alle comunità linguistiche indigene di contribuire direttamente.

1. Introduction

A chain of dependencies links UNESCO goals regarding the Indigenous Communities directly to Language Technologies and then in turn to Language Resources highlighting the critical need for such resources and for innovations that supplement current methods in order to accelerate the production to the benefit of indigenous communities.

1.1 UNESCO Goals

UNESCO supports indigenous languages in order to preserve unique knowledge systems, to promote peace through international cooperation and sustainable development, ensure fundamental human rights, improve education and move toward an inclusive society that acknowledges the value of cultural diversity and heritage[1].

The UNESCO International Year of Indigenous Languages 2019 operates in five key areas of which the following are relevant to Language Technologies and Language Resources: increasing understanding and international cooperation, supporting knowledge sharing relative to indigenous languages", integrating indigenous languages into the digital society, empowering indigenous communities through capacity building, elaborating new knowledge to support growth and development.

The International Conference *Language Technologies for All* (LT4All) promotes the "*human rights and fundamental freedoms of all language users to access information and knowledge in languages that are best understood*" and "*linguistic diversity, truly multilingual internet and language technologies, with special focus on indigenous languages*."[2]

1.2 Language Technologies

The term *Human Language Technology* in the current context refers to any technology that operates upon any human language whether spoken or written or signed. These include Language Identification technologies that recognize what language is being spoken based upon a few seconds of speech. Variants of such technologies also recognize which dialect of a language is spoken and recent work focuses on trying the detect the linguistic origin of the speakers based upon their speech in their native or non-native languages. Speaker Recognition and allied technologies indicate whether two utterances were produced by the same speaker or sort the utterances of a long recording according to the speakers. Speech Recognition technologies automatically produce a transcript of spoken language; related technologies within Dialog Systems convert spoken commands into operations a system can perform. Speech Synthesis or Text to Speech technologies reverse the process and produce utterances of a text or an representation of knowledge retrieved by a system, for example the current time or weather. Information Retrieval and related technologies find written and spoken documents related to a query whether it is expressed as a sequence of search terms or an example document. Information Extraction, and allied technologies find the entities and events and relations among them in a text or spoken document for purposes of answering questions and building knowledge bases. Other Natural Language Processing technologies map the relationship between grammar and meaning in spoken or written language. Finally, Machine Translation, including Speech to Speech Translation, translates content from one language to another.

Together these Language Technologies address many of the goals UNESCO has identified with respect to its work with Indigenous Communities. Specifically they offer methods for constructing and easing access to knowledge in indigenous languages and making that information available to other communities. Similarly language technologies offer ways to provide the world's knowledge to indigenous communities in their own languages. By enabling the development and free flow of information in the languages spoken by user communities, including

[1] https://en.iyil2019.org/

[2] https://en.unesco.org/LT4ALL

indigenous communities, language technologies support the UNESCO goals of protecting fundamental human rights, improving education, promoting sustainable development and international cooperation while acknowledging the value of diversity.

The dominant paradigm in Human Language Technology research and development over the past decades, and the one that has led to such marked advancement, is that of *machine learning*. Under this approach a class of general purpose algorithms develop their ability to process linguistic data by emulating specific human behaviors encoded in annotated data. Thus a system to translate from, for example, Xhosa to Sotho is built from general purpose algorithms and many examples of utterances translated in that way. A great advantage of this approach is that roughly the system can be trained to perform translations between a different pair of languages by providing it with the appropriate data. Another advantage is that system performance tends to increase with the quantity and quality of the "training data" provided and such data has other uses. A disadvantages is that such algorithms tend to require large amounts of training data.

1.3 The Role of Language Resources

Nearly all human language technologies, most modern research into language and great deal of pedagogical materials development rely upon the existence of *Language Resources* by which we mean here: organized collections of records of spoken, and in some cases written, language with annotations that are typically contributed by humans, often aided by technology, in order to support analysis. For purposes of documenting a language the most important of these are what are often called *raw data*: collections of speech, captured in audio or increasingly video recordings, and of texts if the language is written (Good 2011).

Annotation, by which we mean the application of human judgement to raw data, whether directly or mediated by computer, vastly increases its usability (Cieri 2015). The commonest annotation is transcription which generally employs the language's native orthography if one exists in common usage. For purposes of language documentation, descriptive resources such as dictionaries and grammars supplement the raw data. Human language technology developers similarly require raw data but, with a few notable exceptions, tend to rely more heavily on annotations of raw data from which they can extract statistical information than on descriptive resources of the kind created by documentary linguists.

Human language technology performance can be sensitive to the situations under which training data is collected, for example the microphones used, the interlocutors or the genre, thus increasing requirements on quantity, diversity and quality control in language resource development.

2. Language Resource Scarcity

Having traced the chain of dependency from UNESCO goals to language technologies and from language technologies to language resources, we come to the central problem, that of language resource scarcity.

Despite the energetic efforts of a large number of:

- data centers such as Linguistic Data Consortium (LDC)[3], European Language Resources Association (ELRA)[4], Chinese LDC[5], LDC for Indian Languages[6] and the South African Centre for Digital Language Resources (SADiLaR)[7]
- national or regional corpus efforts such as those for Austrian German[8], British English[9], Croatian, Czech[10], German[11], Hungarian[12], Irish (Uí Dhonnchadha 2012), Maltese[13], Dutch[14], Polish[15], Russian[16], Slovakian[17], South Tyrolean[18], Swiss German[19], US English[20], and Welsh[21]
- multination projects to create and share language resources such as CLARIN[22] and META-SHARE[23]
- countless research laboratories that produce language corpora such as LIMSI[24]

it remains true that the number of publicly available language resources is only a tiny fraction of those needed to document and support the development of technologies for the world's languages. Why should this be so?

First, the number of languages in the world is large, more than 7000 by some counts (Eberhard, Simons & Fennig 2019) and the number of resources needed to create a minimal set of technologies for any one language is also not small, perhaps two dozen (Krauwer 1998, Binnenpoorte, et al. 2002, Krauwer 2003). The result is that all of the world's languages lack at least some of the language resources needed, even the languages of the wealthiest nations in the European Union (Rehm and Uszkoreit 2012). However it is also the case that new production does not proceed in a way that maximizes coverage of languages and resource types; rather considerable effort is devoted toward increasing the size of existing resources or producing new versions. (Cieri 2017). Even programs that focus on *under-resourced languages* tend to select from among these language with large

[3] https://www.ldc.upenn.edu
[4] http://www.elra.info
[5] http://www.chineseldc.org
[6] http://www.ldcil.org
[7] https://www.sadilar.org
[8] http://www.aac.ac.at
[9] http://www.natcorp.ox.ac.uk
[10] https://www.korpus.cz
[11] https://www1.ids-mannheim.de/s/corpus-linguistics/projects/corpus-development.html?L=1
[12] http://corpus.nytud.hu/mnsz
[13] http://mlrs.research.um.edu.mt
[14] http://lands.let.ru.nl/cgn
[15] http://nkjp.pl
[16] http://www.ruscorpora.ru
[17] https://korpus.sk
[18] http://www.korpus-suedtirol.it
[19] https://www.chtk.ch
[20] http://www.anc.org
[21] http://codah.swan.ac.uk/?p=334
[22] https://www.clarin.eu
[23] http://www.meta-share.org
[24] https://www.limsi.fr/fr/plateformes-et-ressources/corpus

numbers of native speakers who control large portions of the world's wealth (Cieri 2016).

In short, our current approaches to creating language resources that enable language technology development will not adequately address the scarcity problem in the foreseeable future leaving us to face decades of the same kind of imbalance we currently seek to correct.

3. Innovative Solutions to Scarcity

One reason for the insufficiency of current approaches to creating resources to document the world's language is that it applies a finite and relatively small resource, funding, to a problem that, if not infinite, is at least multiple orders of magnitude larger.

An alternative is to identify renewable sources of the time and intellectual investment required. We take as our model a number of activities that show that the human drive for challenge, advancement, entertainment and the opportunity for people to contribute to their own betterment and that of their local communities and broader society are effectively boundless. This has been made clear repeatedly in the vast numbers of hours spent each day around the world in social media. More immediately relevant, tens of millions of language identification judgements were proffered by players of now defunct GreatLanguageGame (Skirgård, Roberts, & Yencken 2017) and hundreds of millions of contributions have been submitted by nearly two million contributors to the Zooniverse[25] citizen science portal. By providing similar incentives, we offer indigenous language communities a platform in which they can contribute directly to the documentation of their languages and the development of technologies that advance the UNESCO goals sketched above.

LanguageARC is a portal for the Citizen Science of Language, hereafter Citizen Linguistics.

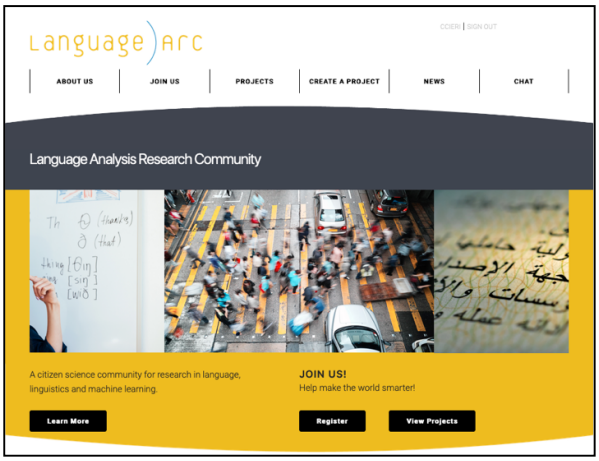


Figure 2: LanguageARC Citizen Linguist portal

LanguageARC presents Citizen Linguists with multiple *projects* to which they can contribute. Each project contains one or more *tasks*; each task involves a simple activity that may be applied to one or more *items*. For example a project might seek to document the linguistic diversity of Italy including the regional and local dialects that are being displaced by the standard language and that have been suppressed by former governments. One task might ask contributors to name culturally relevant items from pictures while another might ask them to describe the people, things and activities they see in a sequence of silent videos. In these cases, the *items* are the pictures and video.

LanguageARC introduces each project via its title, call to action, image and pitch (as in elevator pitch). Each project may also include picture and descriptions of its research team and badges representing its partner organizations. To support community building with the project, each may also offer a range of discussion forums and a blog (currently external). Each task within a LanguageARC project may have its own title, call to action and image as well as a tutorial and reference guide and one additional discussion forum specific to the task.

LanguageARC tasks ask contributors to display provided texts or images or to play audio or video clips and to respond to instructions that are specific to the task or that vary with each item by speaking, enters a text response or selecting one or more items from a multiple choice list.

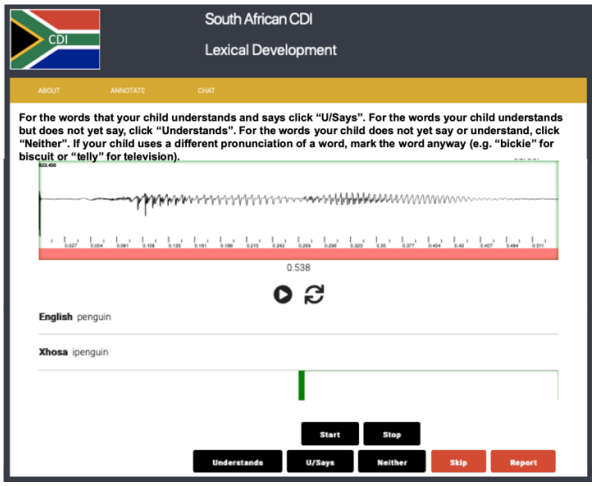


Figure 1: Task prototype for collecting the Communicative Development Inventory in South Africa

Figure 1 contains an image of a prototype, somewhat over-designed to show all possible approaches to collecting the Communicative Development Inventory (Hamilton, Plunkett, & Schafer 2000). for the languages of South Africa. In this case, mothers with young children indicate whether their child has active or passive knowledge of a number of common words for culturally relevant objects. The instructions are in English for the current readership. Beneath the instructions is an audio widget that plays the word in this case in the Xhosa language. Below that are text boxes showing the word in English and Xhosa (presumably only the latter would be used and only if the mother were literate in the language). Beneath that is a recording widget so that the mother can provide a spoken answer and at the bottom of the screen are multiple choice buttons in black and additional red buttons that allow the contributor to Skip

[25] https://www.zooniverse.org

or Report that something is wrong with the item (e.g. the audio did not play).

LanguageARC was built upon a toolkit that the Linguistic Data Consortium has used to create millions of annotations across more than 100 language resource projects over the past decade. The toolkit has been extended to make it open source, portable to new environments and capable even of being deployed to a laptop and taken into the field where internet access is not available. LanguageARC includes a project builder that allows users with no software development experience to create and deploy tasks in less than one hour each assuming the data and instructions are already available in an appropriate format

4. Conclusion

UNESCO aim related to Indigenous Community rely necessarily upon Language Technologies which rely in turn upon Language Resources which are absent for most of the world languages. Current approaches will not solve the language resource scarcity problem in an acceptable timeframe. The use of novel incentives such as those offered by the LanguageARC citizen linguistics portal empowers indigenous communities participate directly in the creation of language resources that benefit themselves principally by enabling technology development but also by encouraging linguistic research and the creation of pedagogical materials.

5. Acknowledgements

The authors acknowledge the support of the National Science Foundation via CISE Research Infrastructure (CRI) grants CRI CI-P 1629923 and CRI CI-NEW 1730377 as well as number partners who make this work possible.

6. Bibliographical References

Binnenpoorte, Diana, Catia Cucchiarini, Elisabeth D'Halleweyn, Janienke Sturm and Folkert de Vriend (2002) Towards a roadmap for Human Language Technologies: Dutch-Flemish experience in Proceedings of the workshop "Towards a Roadmap for Multimodal Language Resources and Evaluation" at LREC 2002, Las Palmas, Canary Islands, June.

Cieri, C. (2017) Addressing the Language Resource Gap through Alternative Incentives, Workforces and Workflows, Keynote Speech at the 8th Language & Technology Conference, November 17-19, Poznań, Poland.

Cieri, Christopher, Mike Maxwell, Stephanie Strassel, Jennifer Tracey (2016) Selection Criteria for Low Resource Language Programs in Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), May 23-28, Portorož, Slovenia.

Cieri, C. (2015) Data Bases and Statistical Systems: Linguistics In James Wright, ed. International Encyclopedia of Social & Behavioral Science 2nd Edition, Elsevier.

Eberhard, David M., Gary F. Simons, and Charles D. Fennig (eds.). 2019. Ethnologue: Languages of the World. Twenty-second edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.

Good, J, (2011) Data and language documentation. In Peter Austin and Julia Sallabank (eds.), Handbook of Endangered Languages. Cambridge: Cambridge University Press. 212–234.

Hamilton, A., Plunkett, K., & Schafer, G., (2000). Infant vocabulary development assessed with a British Communicative Development Inventory: Lower scores in the UK than the USA. Journal of Child Language, 27, 689-705.

Uí Dhonnchadha, E., Frenda, A., Vaughan, B., (2012) Issues in Designing a Corpus of Spoken Irish, LREC: SALTMIL-AfLaT Workshop on "Language technology for normalisation of less-resourced languages, Istanbul, May 2012, edited by G. De Pauw, G-M de Schryver, M. Forcadea, K. Sarasola, F. Tyers, P. Waiganjo Wagach , 2012, pp1-6.

Krauwer, Steven (1998) ELSNET and ELRA: Common past, common future, ELRA Newsletter, Vol. 3:2, May.

Krauwer, Steven (2003) The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap, in International Workshop Speech and Computer (SPECOM-2003).

Rehm, Georg and Hans Uszkoreit, eds. (2012) META-NET White Paper Series: Europe's Languages in the Digital Age, URL: www.meta-net.eu/whitepapers.

Skirgård, H., S.G. Roberts, L. Yencken (2017) Why are some languages confused for others? Investigating data from the Great Language Game, *PLOS ONE*, 12 (4) (2017), p. e0165934, 10.1371/journal.pone.0165934

Tadić, M. (2002). Building the Croatian national corpus. In Proceedings of LREC'2002 (pp. 441–446).

Proceedings, pages 284–287
Paris, UNESCO Headquarters, 5-6 December, 2019.

Can a robot help save an endangered language?

Maximiliano DURAN
ELLIAD , Université de Franche Conté
Besançon, France
duran_maximiliano@yahoo.com

Abstract

A robot using artificial intelligence, a comprehensive set of linguistic resources and pedagogical functionalities may help to preserve Quechua. It can help in M.T of school texts and general culture documentation into Quechua. Written documentation, is essential to keep this language alive. I have been working on such a robot, for several years. I named it Yachaj/expert. The first stage of this project has the following functions: Automatic conjugation, lexical queries of Quechua-FR-SP; elementary spelling checking; and transliteration (alpha version) of texts written in the official spelling of Cuzco, Ecuador or Bolivia to that of Ayacucho and vice-versa.

Keywords: robot, quechua, Yachaj, automatic conjugation, transliteration

Runasimipi

Ama runasimi wañunampaj, allin qispichisqa, allin yachachisqa, llapan rikchaq runasimi cheqap-yachaykunawan, kikin-ruraqqa yanapakuwanchikmanmi runasimi unanchaypi. Chaymi ñuqa, kay ñawpaq qanchis watakunapi runasimita huk kikin-ruraqta "yachachichkani". Paymi yanapakullanman runasimipi, tukuy niraj yachaykunata, yachay-wasikunapi yachachiyta. Paytaqmi, yanapawananchik Fransespi, Castellanopi qellqakunata runasimiman tikrayta, chayna achkallaña runasimipi qellqasqa taqekuna kanampaq. Cheqap-yachaymanta qellqakuna, willakuy-yachaymanta, llimpi-taki-yachaykuna qellqakuna achkallaña runasimipi qellqasqa taqekuna rikurinampaq. Cuzco qelljqa qellajasjata ayakuchu qelqaman tikraypipas yanapawasunchik, kutiriynintapas.

1. A lack of compulsory quechua schooling

The rapid erosion of the Quechua language in just two generations (Fig. 1), in Peruvian territory, confirms its endangered language status. This relative decrease in population objectively shows the danger of Quechua becoming a dead language in the next two generations.

History shows that Quechua has been the victim of many injustices of all kinds that have brought it to this tragic state: historical, administrative, social, linguistic and even psychological.

In Peru, there has been no State policy of compulsory schooling in Quechua in these regions. One consequence of this is the very large and traumatic negative social impact on this population. Since it is not obligatory to speak, and eventually to write correctly in QU, parents are inhibited from transmitting and deepening the knowledge of this language in their children, which inexorably implies the loss of their character of mother-paternal tongue. The language becomes marginalized. The lexical, morpho-syntactic level of the discourse is gradually impoverished. Socially, the child, and subsequently the adult Quechua-speaker, receives from his peers a discriminatory treatment and negative assessment from the monolingual Spanish-speaking speakers, instead of his bilingual knowledge (Quechua-Spanish) being positively weighted. Another consequence is that the incentive for the natural development of the language in the different fields of culture in the society is lost. Cultural production in QU becomes almost a clandestine activity and that is why very few people create literature or songs in this language nowadays; and to aggravate the tragedy, the few who still do, use a lexicon with many "loans" of the language of contact favoring the "quechuallano" or the "medio-lengua" (half quechua-half Spanish).

1.1 Quechua speaking population in Peru

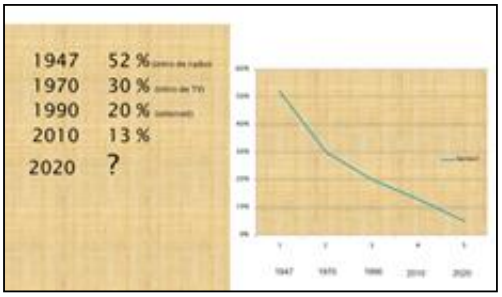


Figure 1: Fall of Quechua speaking population.

1.2 Other converging factors

The other factors that makes worse this erosion are:
a. Incoherent spelling
b. Scarcity of general cultural texts written in QU
c. Social discrimination and racism against quechua speakers

2. Diversification of quechua spelling

In Peru, several spellings have been established by governmental decrees for the QU: one for the Cuzco region, others for Ayacucho and Ancash, among others. In Ecuador, Bolivia and Qrgentina (Santiago del Estero) have also decreed other orthographies. Thus, for the same QU word like language /*simi* there are the following spellings: shimi, simi, ŝimi, čimi. It is as if the various transcripts of

Castilian were made official orthography in each region. That would give rise to hundreds of official spellings of the same Spanish language. In Puerto Rico, for example, one would have to write /Puelto Rico/, following the local pronunciation of /puerto/; in Argentina, one would have to write officially /chó canto/, to spell the conjugated form of "to sing" at the first singular person /yo canto/; or in Chile you would have to write /si pô/ for /si pués/. Historically, of the dozens of transcriptions proposed before the Spanish golden century of literature, the transcription of the speech of the region of Castile was adopted, and since then it is called Castilian.

This diversification of official orthographies for each Quechua dialect region, rather than unifying them, tends to reinforce the socio-linguistic separations. Can we imagine Castilian written in hundreds of official spellings in the world? If this is not possible for that language, why should it be acceptable or good for Quechua?

Can a Quechua expert robot help save the language? A robot, expert in Quechua, using artificial intelligence, and a comprehensive set of linguistic and pedagogical resources, can be an effective aid in the task of safeguarding the language. Such a robot can perform elementary pedagogical tasks, serve for lexicographic mono and multi-language consultation. And, to respond to the lack of written documentation in quechua, the robot can significantly help in the automatic translation of texts written in French or Spanish into Quechua. With this help, in a relatively short time, we could obtain scientific, technical, historical or literary documentation written in this language, which I think is an essential pillar to keep the Quechua language alive (of course there will then be the hard stage of reviewing these translations).

In view of the real shortage of teachers trained for teaching with Quechua as the main language of communication, and the almost non-existent Quechua teachers specializing in technical subjects, scientific, literary or other areas of universal culture, the robot that I am preparing to be an expert on Quechua, whom I have called Yachaj/ *expert*, and who is being programmed to give lessons in QU to could be a valuable aid for the educational system in the Andes. The current technological progress in computing and robotics is a stimulus for the creation of Yachaj and allows us to hope that in a relatively short time it can be equipped with many functionalities, linguistic and more pedagogical resources.

Artificial intelligence (with its various components such as expert systems, the technique of machine learning through artificial neural networks, voice and graphs recognition techniques) applied to QU give us the hope of being able to gestate such an expert robot.

After that, the cloning of the robot Yachaj, for its implantation in schools and other centers and teaching, can be done without great difficulty.

3. The Quechua expert robot Yachaj

For more than 7 years I have been working on the development of such a robot. Several of the basic linguistic resources, such as electronic dictionaries, have been prepared since 1990. In this work, I count with the scientific council of researchers of the Laboratory of Linguistics and Informatics of the University of Zurich, of Grenoble and Besançon and some colleagues of INALCO of Paris At the current stage of its development, Baby robot QU, is able to perform, in laboratory, the following linguistic and pedagogical functions.

3.1 Lexicography

The Yachaj robot have around twenty bilingual electronic dictionaries such as:

. Dictionary DG- SP-QU containing more than 43.000 QU-SP entries and 34.000 QU-SP ones.

. Dictionary DG FR-QU containing around 36.500 entries and DG QU-FR containing around 21.200 entries.

. Dictionary of Quechua conjugated forms (more than 2 million forms) with their respective translations FR-QU and QU-FR.

. MWU Dictionary, containing multi-word linguistic units in QU-FR and QU-SP,

. LVF_QU Dictionary, containing 8.600 FR-QU verbs, from the Lexique de Verbes Français de Dubois & Dubois-Charlier (1997), translated by M. Duran (2013),

. An original Quechua scientific-technical lexicon, etc.

Figure 2: QU-SP general dictionary (38 000)

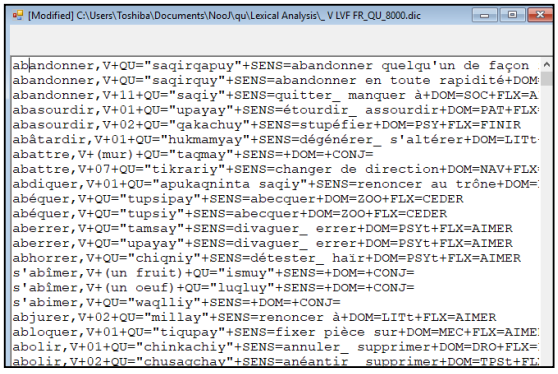


Figure 3: LVF FR-QU verbs (8 600)

3.2 Morphology

Yachaj "knows" how to apply thousands of rules of QU morphology, and he knows how to apply them. It knows exhaustively all the nominal suffixes SUF_ N, adjective suffixes Suf_A, and verb suffixes inter and post positional IPS and PPS, and knows how to make the grammatically valid combinations between two or more suffixes. It knows all the semantic values of about 240 language suffixes. In the following figure we show the extracts of the inflections (655) of a noun like wasi/ house and the verbal forms (7743) that Yachaj obtains by applying conjugation rules, of derivation with 1-3 suffixes.

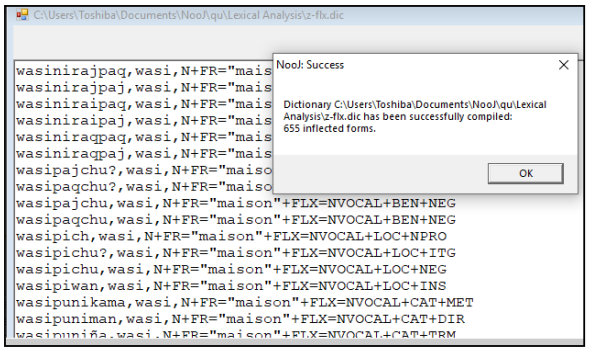


Figure 4: Inflectional forms of wasi (1-3 suf.) (655)

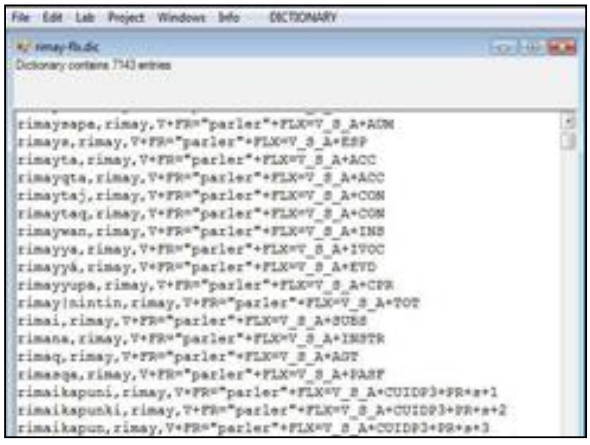


Figure 5: Inflected forms of verb rimay (1-3 suf.) (7743)

4. Spelling checker

The robot contains a user interface to get help in spelling correction of texts written in official QU of Ayacucho. Work is under way on spell-checkers for the other dialectal versions

5. Automatic inter-dialectal transliterator

Yachaj is able to carry out the inter-dialectal transliteration between the Cuzco>Ayacucho versions and vice versa, in the laboratory. Following the same technology, the programming of automatic transliteration has been initiated to help pass from an official spelling, Cuzco, Ancash, Ecuador, Santiago del Estero or Boliviano to Ayacuchano and vice versa. This system can be used to solve this writing controversy of "Into which written quechua? and allow us to overcome that divisive barrier that each idiosyncrasy claims to make prevail the quechua and its writing of 'its" region is correct and the rest is not. If a Cuzco way of writing is not readable to the Ayacucho-Quechua-speaking, by pressing the "transliterar" button to the Ayacuchano, he will be able to get his text in official Ayacucho spelling. And vice versa.

Here is an excerpt from the automatic transliteration of a poem by Jayme Araoz Chacón (2008) written in official Cuzco spelling into Ayacucho spelling.

Cuzco spell	Ayacucho sp.	Castilian
Rumi	Rumi	*Piedra*
Waklaw chimpaman	Waklaw chimpaman	*La piedra que al*
Rumi chanqasqay	Rumi chamqasqay	*frente lancé*
Waklaw chimpaman	Waklaw chimpaman	*La piedra que al*
Rumi wikch'usqay	Rumi wischusqay	*frente arrojé*
Maytaq kunan	Maytaq kunan	
Rikhurimunchu	Rikurimunchu	*Porqué ahora no*
Maytaq kunan	Maytaq kunan	*aparece?*
Kutimunñachu	Kutimunñachu	*Porqué hoy ya no*
Chhaynallataqsi	Chaynallataqsi	*regresa?*
Warma munasqay	Warma munasqay	
Chhaynallataqsi	Chaynallataqsi	*De modo parecido*
Yana wayllusqay	Yana wayllusqay	*La mujer que quiero,*
Maytaq kunan	Maytaq kunan	*Del mismo modo mi*
Kutimunñachu	Kutimunñachu	*amada ves cómo ya*
Maytaq kunan	Maytaq kunan	*no vuelve, ves cómo*
rikhurimunchu	rikurimunchu	*hoy ya no se aparece*

6. Conclusion

We have presented the first stage of a quechua expert robot. To build it, we are using an important number of linguistic resources obtained mainly with the aid of the linguistic platform NooJ (Silberztein 2003, 2015). We are using also some machine learning techniques. For the time being it is capable to perform a certain number of pedagogical functionalities which may be useful for elementary school learning in quechua. The first stage of this project have the following functions: Automatic conjugation, lexical queries of Quechua-FR-SP; elementary spelling checking; and an alpha version of a transliterator of texts written in official spelling of Cuzco, into Ayacucho spelling and vice-versa.

7. Bibliographical References

Araoz, J. (2008). Llaqta Taki Harawi. *UNICEF. Danny's EIRL. Cuzco Peru*,

Dubois, J. et Dubois-Charlier F. (1997). *Les verbes français*, Larousse, Paris.

Duran, M. (2009), *Dictionnaire Quechua-Français-Quechua*, Editions HC. Paris.

Duran, M (2013), *Formalizing Quechua Noun Inflexion*. Formalizing Natural Languages with NooJ. Edited by A. Donabédian, V. Khurshudian and M. Silbeztein. Cambridge scholars. Newcastle upon Tyne,.

Duran, M.(2013) *Formalizing Quechua verbs Inflexion*. Proceedings of the NooJ 2013 International Conference, Saarbrücken,.

Silberztein M (2015). *La formalisation des langues*. ISTE Editions. London.

Silberztein M(2003). *NooJ Manual*. http:/www.nooj4nlp.net (220 pages, updated regularly).

Proceedings, pages 288–290
Paris, UNESCO Headquarters, 5-6 December, 2019.

Accessing and Understanding contents in Portuguese by foreigners in scientific digital libraries: can this methodology be generalized to other languages?

Cláudio MENEZES
University of Brasília
Department of Foreign Languages and Translation (LET)
*Cam*pus Universitário Darcy Ribeiro
70919-900 Brasília, DF, Brasil
claudiomenezes@unb.br

Abstract

In digital libraries, remote access to documents has become frequent. Some examples: http://www.ndltd.org/ and http: //bdtd.ibict.br/vufind/ in Brazil which allow access to the text of original documents. Since the number of foreign students in universities has increased, there is a need for a service for them. However, the full translation of these documents would be a herculean task. This research identified some obstacles encountered by foreigners to access and understand scientific content and offers a methodology supported by a computer application facilitating its understanding by Francophone students. It can be adapted to any language pairs, including sign languages and braille.

Keywords: automatic summarization, digital libraries, translation, scientific contents

Résumé

Em bibliotecas digitais, o acesso remoto a documentos tem sido a regra. Alguns exemplos: http://www.ndltd.org/ and http: //bdtd.ibict.br/vufind/ no Brasil permitindo acesso ao texto dos documentos originais. Como o número de estudantes estrangeiros nas universidades tem crescido, há a necessidade de lhes oferecer um serviço específico. No entanto, a tradução completa desses documentos seria uma tarefa hercúlea. Esta pesquisa identificou alguns obstáculos encontrados por estrangeiros para acessar e compreender conteúdo científico e oferece uma metodologia baseada em uma aplicação computacional facilitando a compreensão por estudantes francófonos. Pode ser adaptada para quaisquer pares de línguas, incluindo a língua de sinais e o braile.

1. Introduction

This paper presents a research aimed to identify the obstacles encountered by foreigners to access and understand scientific content in Ph. D. thesis and M. Sc. Dissertations in digital libraries and shows a methodology supported by a computer application that could improve their understanding by Francophone students. The proposed methodology can be adapted to any language pairs. It can also be considered for adaptation to sign languages, braille and oral communication.

2. Research synthesis

The identification of barriers to access and comprehension of scientific texts by French-speaking students at the University of Brasilia and the University of Lille 3 (Charles de Gaulle University) was collected through questionnaires[1], which led to to obtain data about linguistic expertise and knowledge and use of automatic language processing software.

As a general outcome, two situations have been identified: 1) the simple withdrawal of the use of scientific bibliography in Portuguese by the foreign user due to insufficient knowledge of the Portuguese language; 2) lack of knowledge of natural language processing tools to facilitate access to and understanding of scientific texts available in digital libraries with texts written in Portuguese.

Based on this observation, we have begun an exhaustive search of computer tools that can help the foreign user to access and understand scientific contents available in a *corpus* of theses and dissertations of the University of Brasília, part of the Digital Library of Theses and Dissertations (BDTD), project coordinated by the IBICT (Brazilian Institute of Information Science and Technology).

3. Description of the methodology

The key idea of the proposal is to build a semantic representation of scientific texts in order to avoid the need for a complete translation of original theses and dissertations. To do this, four computing technologies were used: filtering, automatic summarization, machine translation and sentence alignment, as further explained.

Since the documents available in digital libraries are generally in **pdf** format and contain chapters or sections without major semantic interest, the first computing resource used is a **filter**. This feature produces a new document in **txt** format, debugging sections without semantic interest (acknowledgements, presentation, index, bibliography for example). This new document contains only the chapters of the thesis or dissertation to be studied. The conversion of the format to .txt is due to the need to use programs in which input files are required in this format.

[1] The pre-test was conducted with French-speaking students from the Portuguese Teaching and Research Center (NEPPE), UnB. A second data collection was carried out with students in Information Sciences at Charles de Gaulle University (Lille 3)

The second component of the computer tool - the **automatic summarizer** - makes it possible to produce a smaller text formed by the most relevant sentences of the original document. There are several techniques and criteria for creating relevant abstracts. To provide greater flexibility to the user, the input parameters are the start and end page numbers of the text to be summarized. However, it is recommended to choose as parameters the first page of the first chapter and the last page of the last chapter of a thesis. The user must also provide the desired **compression rate**, which will indicate the size of the summary to be created and translated. Two automatic Portuguese automatic summarizers were tested: GISTSUMM and GENSIM. Note, however, that in one of the experiments performed, accuracy and coverage are calculated using the "ROUGE" (Recall-Oriented Understudy for Gisting Evaluation)[2] program to get an idea of the quality of the summary produced by GISTSUMM.

The third component of the application, **machine translation**, provides the translated text in the target language. In our case, we work with the language pair (FR, PT), but the chosen software has the ability to translate into six languages: Portuguese, French, Spanish, German, English and Japanese.

The fourth component – **paragraph alignment** - produces a parallel text (bi-text) in Portuguese and in the target language. In the targeted research, experiments were carried out with the PT - FR pair.

The flow diagram available at https://github.com/leandro2r/automatic_summarizer illustrates the relationships between each component of the IT tool used, accessible at http://multilingua.cdtc.unb.br:8080/

4. Extension of the methodology to other linguistic pairs

Current research was conducted with Francophone students. However, the four components of the proposed methodology can be adapted to other language pairs. Naturally, it will be necessary to use filters, automatic summarizers, translation programs and sentence alignment programs able to work with the chosen language pair.

The methodology adaptation to assistive technologies such as sign languages and text-to-voice technologies is also a development to be further explored.

The extension of the proposed methodology to other language pairs is therefore an evolutionary work that can be developed in other similar projects.

5. Conclusion

Whether in libraries or directly by its users via the Web, the use of natural language processing tools to enable access and understanding of content in another language is still very embryonic. It is hoped that the methodological proposal for this research will be adopted and improved by other research groups interested in the subject, with the aim of broadening multilingualism in cyberspace and promoting the linguistic vitality of a greater number of languages. in the digital world.

6. Bibliographical References

BOJAR, Ondrej et al. Findings of the 2016 conference on machine translation (wmt16). **Proceedings of the First Conference on Machine Translation**, v. 2: Shared Task Papers, p. 131–198, Berlin, Germany, aug. 11-12, 2016.

BRANCO, António et al. **The Portuguese Language in the Digital Age**. Berlin: Springer, 2012.

FRIAS-MARTINEZ, E. et al. Automated user modeling for personalized digital libraries. **International Journal of Information Management**, v. 26, n. 3, p. 234-248, 2006.

GALE, William A.; CHURCH, Kenneth W. A program for aligning sentences in bilingual corpora. **Computational linguistics**, v. 19, n. 1, p. 75-102, 1993.

GUPTA, Vishal; LEHAL, Gurpreet Singh. A survey of text summarization extractive techniques. **Journal of emerging technologies in web intelligence**, v. 2, n. 3, p. 258-268, 2010.

LLORET, Elena et al. Compendium: a text summarisation tool for generating summaries of multiple purposes, domains, and genres. **Natural Language Engineering**, v. 19, n. 2, p. 147-186, 2013.

MARCONDES, Carlos H. et al (Org.). **Bibliotecas digitais**: saberes e práticas. 2. ed. Salvador: Ufba, 2006.

MÁRDERO, Arellano. Àngel. Serviços de referência virtual. **Ciencia da Informação**, Brasilia, v. 30, p.1-15, 2001.

MENEZES, Cláudio; BAPTISTA, Dulce Maria. Metodologia de Acesso a Dissertações de Mestrado de Tradução por Estrangeiros: Uma abordagem preliminar.
Revista Iberoamericana de Ciência da Informação, Brasilia, v.10, n.1, p. 154-163, jan./jul. 2017. Disponível em http://periodicos.unb.br/index.php/RICI/article/view/16462/18074. Acesso em 16.10.2017

MENEZES, Francisco Cláudio Sampaio de. O Multilinguismo e as Novas Tecnologias das Línguas no Século XXI. **Belas Infiéis**, Brasilia, v. 4, n. 12015, p.85-98, 01 jun. 2015. Disponível em: <http://periodicos.unb.br/index.php/belasinfieis/issue/view/1175/showToc>. Acesso em: 15 nov. 2015.

MIHALCEA, R.; TARAU, P. TextRank: Bringing order into texts. Association for Computational Linguistics. **EECS News**, jul. 2004. Disponível em: https://web.eecs.umich.edu/~mihalcea/papers/mihalcea.emnlp04.pdf. Acesso em: 19 jun. 2017.

RINO, Lúcia Helena Machado et al. Summarizers of Texts in Brazilian Portuguese: Lecture Notes on Artificial Intelligence. In: 17TH BRAZILIAN SYMPOSIUM ON ARTIFICIAL INTELLIGENCE, 1., 2004, São Luis. **Proceedings of the 17th Brazilian Symposium on Artificial Intelligence.** São Luís: Springer-verlag, 2004. v. 1, p.

235 - 244. Disponível em: <https://www.researchgate.net/publication/220974768_A_Comparison_of_Automatic_Summarizers_of_Texts_in_Brazilian_Portuguese>. Acesso em: 29 set. 2004.

SOUZA, C.F.R.; NUNES, M.G.V. **Avaliação de Algoritmos de Sumarização Extrativa de Textos em Português**. Relatórios Técnicos do ICMC-USP. NILC-TR-01-09, Novembro 2001

UNESCO, A Decade in Promoting Multilingualism in Cyberspace", Disponível em http://unesdoc.unesco.org/images/0023/002327/232743e.pdf, Acesso em: 05 out. 2017

Proceedings, pages 291–295
Paris, UNESCO Headquarters, 5-6 December, 2019.

Challenges and Opportunities in Processing Low Resource Languages: A study on Persian

Mehrnoush Shamsfard
NLP Research Lab, Shahid Beheshti University, Tehran, Iran
m-shams@sbu.ac.ir

Abstract

This paper discusses the importance of language processing and its challenges. It first defines low resource languages and their influencing factors. Then talking about the Persian Language, discuss the situation of Persian in this field of study. Following the discussion, some major Persian language resources are introduced. and describing available methods, some challenges and opportunities are discussed. At last the conclusion section suggests some steps for moving Persian from a low resource language toward a high-resource one, including using cross-lingual embeddings, standardization of test data, running various challenges on Persian data and encouraging startups to build their business in this field.

.**Keywords:** Language technology, Low resource language, Persian

Résumé

این مقاله به اهمیت پردازش زبان طبیعی و چالشهای آن می پردازد. به این منظور ابتدا با ارائه تعریف زبانهای با منابع محدود، عوامل تاثیرگذار در این محدودیت را بررسی می'نماید. سپس با بحث در مورد زبان فارسی، به جایگاه آن از نظر محدودیت منابع اشاره می‌نماید. در ادامه برخی از منابع زبانی توسعه‌یافته برای زبان فارسی معرفی شده، کمبودهای این منابع مورد توجه قرار می‌گیرند. مقاله سپس با بیان روشهای فعلی، به بحث در مورد چالشها و فرصتهای موجود در پردازش زبان فارسی پرداخته و در نهایت در بخش نتیجه گیری راهکارهایی برای افزایش منابع زبانی آن ارائه می‌نماید.

1. Introduction

There are 6-7 thousand languages in the world with different size of native speaker population. Some of these languages such as English, Spanish or French are very popular to be learnt and spoken as the second language (L2) while some others may be just spoken or written by a small number of native speakers.

Most of the languages which have orthography and written form have written documents and many of written languages have electronic documents on the web or local media. So being able to process these documents is a must. Resource-rich languages are those with matured language technology, many language resources and processing tools and applications and on the opposite side, resource-poor languages suffer from the lack of data, language resources, language technology and processing tools and applications. Actually a small fraction of languages is resource rich or high resource, among which, English has a specific position; it is actually a laboratory language which has attracted a big share of computational linguistic efforts and researches in the world. Even non-native English speakers work on English language to develop or enhance resources, algorithms, methods, processing tools and applications or create data or text to present their ideas, news, achievements and knowledge in this language (this paper is an evidence of this fact). In this way, some few resource rich languages are used more and more in the digital world and cyberspace and the others are weakened. This forces the speakers of the weakened languages to learn those resource-rich languages to be able to transfer their ideas, present their culture and language, talk about their achievements, and so on. Thus the resources of resource rich languages become richer every day and riches get richer.

On the other hand, the weakened languages will gain lower share of documents in the cyberspace and move to being removed from digital media gradually. Paying attention to these languages and providing essential technologies to enable their speakers to communicate and share their data, experiences and knowledge in their own native language not only helps to save diverse languages but also makes the data, information and knowledge from different cultures, geographic regions, political governments, etc. available to linguists, social studying and aiding communities, weather and environment watchers, politicians, militaries and so on. Thus, enhancing language technologies for low resource languages is beneficial to all.

2. Low Resource languages

There are different definitions for low resource languages. Although low-density or indigenous languages are included in this category, it also includes some languages with a large size of native speaker population. LORELEI defines low resource languages as the languages for which no automated human language technology exists. This definition excludes languages for which there are some tools but they do not cover all aspects of human language or does not have good performance such as Persian. According to Duong (2017) a language is considered low-resource for a given task if there is no algorithm using currently available data to automatically do the task with adequate performance. This definition implies that a language is considered low-resource based on a specific task. In this definition for example, Persian is not a low-resource language with respect to part-of-speech tagging of formal written text as the performance of a tagger is 96% precision, while it is low resource for POS-tagging of colloquial informal texts or for other tasks such as recognizing multi word expressions.

There are various factors affecting languages; keeping them low resource or helping them to become high-resource. Among these factors the availability of language resources and technology, expertise in natural language processing, financial supports and political issues are the most important. These factors are themselves effective on each other in a cycle.
The most important factor which makes a language low-resource is the technical one; availability of language resources and technology. In the next section we talk about its challenges and opportunities.

The other factor which is highly influenced by the others is human expert availability. Many researchers from all over the world (even low resource language native speakers) are serving language processing technology for resource-rich languages such as English, because

- There are sufficient available English training data for machine learning algorithms and so novel methods can be tested easily.
- There are many tested, reliable tools and toolkits, and many pieces of source codes or reproducible methods available for conducting a research for English while starting the same research for a low resource language is very hard and time- and effort- consuming and may lead to lower performance measures at the end.
- Journals and conferences related to language technology usually ask authors to compare their work with other similar works and as in many cases there is no similar work in the low resource language the researcher has to pay attention to methods in resource-rich ones such as English too.
- As the audiences of a language in minority are much less than the audiences of languages in majority, publishing research paper about minor languages in journals is harder. Many Journals does not accept such papers due to their audiences' needs and even after publication the paper may have fewer citations. This problem lead some researchers to change their work focus from their own native language to the English language.

Two other factors are economics and politics. Countries with rich economy can dedicate sufficient financial resources and funds for research on language technology while the others cannot. The rich countries may even be interested on processing some non-native languages due to some political, economic or cultural issues so their economy may help other language resources to grow.

On the other hand, political issues may affect or prevent the development of language resources and enhancement of language technology. For example, in Iran which is under sanctions, the sanctions are applied to any aspect of Iranian lives including science and technology. Unavailability of powerful processing devices, source files and libraries from code providers such as google, some softwares even compilers and interpreters, data, and even rejecting research papers due to political issues are some of the problems which sanctions have brought to Iranian researchers in this field.
All of these factors and some others make a language as low resource or help it to grow to high resource level.
In the rest of the paper, we study the situation of Persian in this field of study and propose some steps toward providing language resources and technology for Persian

3. The Persian Language

According to traditional classification, Persian with the Indo-Aryan languages constitutes the Indo-Iranian group within the Satem branch of the Indo-European family. This group consists of Persian, Pashto, and Kurdish.
Persian is the official language of Iran, Afghanistan and Tajikistan with more than one hundred million speakers and also is spoken in more than six other countries. According to its geographical position Persian speaker countries are neighbor to Arabic countries and so there are a lot of loaned words entered to Persian from Arabic. Although there are many differences between Persian and Arabic language especially according to grammar and syntactic features, there are some similarities in lexical level and some Arabic derivational rules have come into this language (Shamsfard, 2011).
Although Persian is the official language of Iran, there are some other languages spoken in Iran such as Kurdish, Turkish and Arabic and sometimes documents to be processed are a mixture of these languages. Even for Persian, there are various types and dialects. For example, Persian texts can be written in colloquial or formal Persian. Colloquial texts are used in daily conversations, non-formal short messages (SMS), blogs, social media, emails and some books (especially novels) while formal texts are used in formal conversations, formal or scientific documents, news, educational and many other books. Persian colloquial and formal texts are very different from each other especially in the lexical level. They need different lexicons, different training and testing datasets, and even different grammatical rules for NLP tasks. Most of the language resources and tools are dedicated to formal Persian.
Some linguistic features of the formal Persian language are as following (Shamsfard, 2011);
Persian is written right-to-left. It is a pro-drop language with canonical SOV word order with a lot of frequent exceptions in word order, which have turned Persian to a free word order language. Verbs are marked for tense and aspect and agree with the subject in person and number with some exceptions. Although verb-final, Persian is otherwise mostly head-initial.
Persian letters have one to four forms of writing. Different forms are used depending on the position of the letter within the word which may be initial, medial or final (isolated). There are various scripts for writing Persian texts, differing in the style of writing words, using or elimination of spaces within/between words, using various forms of characters and so on. Persian is a derivational and generative language in which many new words may be built by concatenating words and affixes. Usually none of the short vowels are written in a Persian sentence. So facing homographs and homonyms are popular ambiguities in Persian. Usually there is no definite article in a Persian sentence while most of the nouns appear with one in English. Unlike English there is no female/male distinction for Persian pronouns and there is no rule for appearing uncountable nouns in singular form. Even words which are uncountable may appear in plural form.
In Persian each verb conjugates in its own tense while in English the tense of the other sentence verb must be considered. In Persian words and phrases may be omitted in a sentence according to a syntactic or semantic symmetry. Omitting the subject is also very popular in

Persian sentences. In this case the agreement (person and number) embedded in the verb can play the subject role.
In Persian in many cases adjectives can be inserted in place of nouns without any lexical change and this may cause structural or semantic ambiguities in noun phrases.

Working on Persian language processing is a growing field. The early efforts in this field go back to late 1980s and the very first systems such as Dena for Persian text understanding (Fahimi & Shamsfard, 1995) were introduced in 1990s. Many language resources, tools and applications have been developed for Persian during the last 25 years. But still Persian is far from English in language technology and can be assumed a low resource language in many tasks. In the next section we mention some of the main language resources for Persian but left the survey on tools and applications for another paper due to the small size of this paper.

4. Language Resources for Persian

Available language resources can be divided into the following categories. In each category some are named.

- Corpora: There are various corpora available for Persian, with different sizes and taggings. Peykareh (Bijankhan et al., 2011) with about 8 million tagged tokens, FLDB (Assi, 1997), and Hamshahri (AleAhmad et al., 2009), are the most famous general POS tagged corpora. The large Beheshti corpus of more than 4 billion tokens, dump of Wikipedia[1], blogfa corpus and corpus of tweeter posts are raw corpora used for language modeling and building word embeddings. PAYMA (Shahshahani, et al., 2015), ArmanPerosNERCorpus (Poostchi, et al., 2016), A'laam (Hosseinnejad, et al., 2017) are instances of NER tagged corpora for Persian. Parallel corpora such as Mizan (Kashefi, 2018) with more than 1 million sentence pairs and TEP (Pilevar, et al., 2011) with about 550,000 pairs of movie subtitles, and comparable corpora such as (Hashemi, et al., 2010) with 7500 document pairs are another type of corpora. Task specific corpora such as Mahtab plagiarism detection corpus (Mashhadirajab, et al., 2016) with 20000 documenrts and about 10000 suspecious documents, Behehshti sense tagged corpus (Rouhizadeh et al., 2019), Treebanks such as (Mirzaei and Safari, 2018) for discourse and (Rasooli et al., 2013) for dependency fall in this category too.

- Lexicons and Thesauri: General lexicons such as zaya (Eslami, et al., 2004) with 55000 entries and FarsVajeh (Shamsfard and Jafari, 2017) with about 80000 entries are available. In Farsvajeh, each lexeme is associated with its various written forms and the preferred orthography suggested by APLL, its phonetics, POS tags, frequency in a corpus, and morphological structure (inflectional, derivational, compound, ...). Although there are some larger lists of words, we couldn't find lexicons larger than 100,000 entries with phonetic, morphological, and syntactic information. Of course the electronic version of some Dictionaries such as sokhan, dehkhoda, mo'een, etc. are available (mostly without a legal license) but they are not structured. There are also some general thesauri such as Fararooy (1998) (almost a translation of Ruget's) and some domain specific thesauri such as the ones developed by Irandoc[2] (eg (Norouzi, 2003). Sentiment lexicons such as PerSent (Dashtipour et al., 2016), LexiPers (Sabeti et al., 2016)), HesNegar (Asgarian, et al., 2018) and SentiFars (Dehkharghani, 2019) fall in this category too.

- Wordnets and knowledge graphs: FarsNet is the first, biggest and most reliable wordnet for Persian. The third version of it contains more than 100,000 lexical entries organized in more than 40,000 synsets with glosses, examples and various semantic relations. It is developed semi-automatically and revised manually. Persian wordnet of Tehran (Taghizadeh and Faili, 2018) is another work in this field. It is translated automatically from Princeton WordNet. FerdowsNet as the third Persian wordnet which is smaller than the first two is developed by Ferdowsi University but is not available to public. According to wordnet Persian is among resource-rich languages. FarsBase (Asgari, et al., 2019) is the Persian knowledge graph with 5,582,589 links to external datasets.

- Datasets: the main problem in Persian resources is here. Persian lacks large reliable data sets for training and testing systems for different NLP tasks. Even for tasks that have large known English datasets such as question answering (QA), chatbots, text generation, WSD, multi-word expression (MWE), sentiment analysis, entailments and paraphrases, etc., there is either no dataset or the available datasets are too small or not reliable. Some of the available datasets are wsd data (Rouhizadeh, et al. 2019), Pars-ABSA for aspect based sentiment tagged opinions (Ataei, et al., 2019), and parallel formal and informal Persian language (under construction),

It seems that the open problems are development of corpora with various tags (except POS and NER) and datasets for various application tasks such as those mentioned in the previous paragraph.

5. Methods, Challenges and Opportunities

In recent years, the language technology is shifted from rule based systems to statistical ones and now to deep neural networks and distributed semantics with dense vectors. In the current trend of research, we need a huge amount of data to train deep neural systems and low resource languages are those for which such a data is unavailable. Even for unsupervised methods which do not need training data, at least we need standard, accurate, reliable test datasets with good coverage to test the developed systems. Persian is among the languages which suffer from the shortage of language resources such as tagged corpora and train and test datasets. For instance, datasets of questions and answers, entailment sentences, paraphrase pairs, chats, restyling sentences, metaphors, and texts and their internal representations are some examples which are available for English but not for Persian. While datasets for tasks such as syntax parsing over constituency or dependency (tree banks), sentiment analysis, named entity recognition, word sense disambiguation and machine translation (parallel corpora) are available for Persian but mostly their size, and sometimes their accuracy and quality are lower than the corresponding data in English.
The shortage can be eliminated by either creating resources and tools for the low resource languages from scratch or trying to adapt/use resources and tools in other languages.

[1] https://dumps.wikimedia.org/fawiki/

[2] https://irandoc.ac.ir

Thus we can assume four major approaches to handle the shortcoming of data and resources for Persian language:

- Direct Translation of data from English or any other language which has the data: Unfortunately, this approach is not a good one in many cases, as the quality of translators are not admissible and the translated data should be revised and corrected manually which takes a long time itself. The created dataset is biased to the source language and may omit the linguistic phenomenon of Persian. On the other hand, due to differences of the two languages, the word-by-word alignment may be impossible or some features such as POS tag may be changed during the translation and so the method may be inappropriate for some sorts of datasets or some types of tasks (eg. POS tagging). The small dataset for Persian WSD (Rouhizadeh, et al., 2019) is an example of this method.
- Processed translation of English or other language's data: In this method, in addition to translation, the data will be under some processes to enhance the translated data and remove uncertain parts. In other words, this approach utilizes cross-lingual methods and bilingual (or parallel) resources to build Persian datasets from English ones. This method is more complicated than the previous one but its results are more reliable than translation without human revision. It is faster than the next method and can create larger datasets with lower costs. The method is more suitable for translating words and not texts and has the drawback of biasness same as the previous method. The Persian wordnet developed by Tehran University (Taghizadeh, et al, 2018) is an example of this method.
- Creating data for Persian: Many researchers try this approach. It is time and cost consuming and the results are usually smaller than the corresponding datasets in English. But the dataset is not biased to any other language and it's usually more accurate and precise than the previous ones. FarsNet (Shamsfard, et al. 2010; khalghani & Shamsfard, 2018)), Dataset of Mahtab (Mashhadirajab et al., 2016), Persian treebank (Rasooli, et al, 2013) and ArmanPerosNERCorpus (Poostchi, et al., 2016) are some examples of this method.
- Using cross-lingual and transfer based deep methods to use English or other language's data to perform Persian tasks: Moving toward cross-lingual efforts is a way to use e.g. English embeddings for Persian tasks. It seems that for each deep approach to a problem if the embeddings be cross-lingual then the method will work for Persian as well as the English one using English training set. This method is expected to be faster than the previous ones and can utilize larger, more reliable datasets for Persian tasks.

6. Conclusion and Future Work

In some domains we have enough resources while in some others, the resources are either rare or missing. For example, there are enough POS and NER tagged corpora, good wordnets, huge amount of untagged raw texts, several sentiment lexicons, and medium size parallel (Persian-English) corpora. But still, some fundamental tasks such as tokenization have problem in available corpora and datasets. So compound words and verbs are not assumed as one token and this makes a lot of problems in various tasks especially in word embedding.

We need enhanced corpora, tagged by various features such as entity linking and coreferences and various datasets such as training and testing data for question answering (QA), chatbots, text generation, WSD, multi-word expression, sentiment analysis, entailments and paraphrases, etc.

It seems that (1) moving towards Cross-lingual embeddings, (2) establishing a research center for standardization and generation of test data and creating testbeds and benchmarks to evaluate resources and tools and (3) running various challenges on Persian data such as those ran by SemEval, SenseEval, TREC, … and (4) encouraging and helping startups and companies to build their business in this field, are some possible actions which will speed up the development of Persian resources and move this language from being low-resource.

7. Bibliographical References

AleAhmad, A. Amiri H., Darrudi E., Rahgozar M., and Oroumchian F. (2009). Hamshahri: A standard Persian text collection. *Knowledge-Based Systems Journal*.

Asgari B., M., Hadian A., Minaei-Bidgoli M., (2019). FarsBase: The Persian knowledge graph, *Semantic Web*,

Asgarian E, Kahani M, Sharifi S. HesNegar: Persian Sentiment WordNet (2018). *JSDP*. 15 (1):71-86

Assi, M. (1997). Farsi Linguistic Database (FLDB), *The International Journal of Lexicography*, 10(3):6, Oxford University Press.

Ataei, T.S., Darvishi, K., Minaei-Bidgoli, B., Eetemadi S., (2019). Pars-ABSA:an Aspect-based Sentiment Analysis dataset for Persian, arXiv:1908.01815.

Bijankhan, M. Sheykhzadegan, J. Bahrani, M. and Ghayoomi, M., (2011). "Lessons from Building a Persian Written Corpus: Peykare," *Language Resources and Evaluation,* 45(2):143–164.

Dashtipour, K.,Hussain, A., Zhou, Q., Gelbukh A., Hawalah A.Y.A., Cambria E., (2016). PerSent: A Freely Available Persian Sentiment Lexicon. In BICS 2016, 8th International Conference on Brain-Inspired Cognitive Systems.

Dehkharghani R., (2019). SentiFars: A Persian Polarity Lexicon for Sentiment Analysis, *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP*), 19(2).

Duong, L., (2017). Natural language processing for resource-poor languages, PhD dissertation, University of Melbourne, Australia.

Eslami, M., Sharifi, M., Alizadeh, S., Zandi, T., (2004). 'Persian Generative Lexicon', 1st workshop on Persian Language and Computer, pages 6-11.

Fahimi, M., Shamsfard, M. (1995) Dena: A Persian Text Understanding System, In Computer conference of Computer Society of Iran (CSI).

Fararooy, J., (1998). *Persian Thesaurus*, Ibex Publishers.

Hosseini, P., Ahmadian R. A., Maleki, H., Anvari, M., Mirroshandel, S.A., (2018). SentiPers: A Sentiment Analysis Corpus for Persian, arXiv:1801.07737.

Hosseinnejad S, Shekofteh Y, Emami Azadi T. A'laam Corpus: A Standard Corpus of Named Entity for Persian Language. (2017). *Journal of Signal and Data Processing (JSDP).* 14 (3):127-142.

Kashefi O., (2018). MIZ¯AN: A Large Persian-English Parallel Corpus, arXive

Khalghani, F., Shamsfard, M., (2018). Extraction of Verbal Synsets and Relations for FarsNet. The 9th Global WordNet Conference (GWC 2018).

Mashhadirajab, F., Shamsfard M., Adelkhah R., Shafiee F., and Saedi C. (2016). A Text Alignment Corpus for Persian Plagiarism Detection. In FIRE 2016, pages 184-189.

Mirzaei, A. and Safari p., (2018). Persian Discourse Treebank and Coreference corpus, LREC 2018, pages 4049-4055.

Norouzi, M., (2003) *Engineering Thesaurus*, Irandoc.

Pilevar M.T., Faili H., Pilevar A.H. (2011) TEP: Tehran English-Persian Parallel Corpus. In Computational Linguistics and Intelligent Text Processing. CICLing 2011. Pages 68-79. Springer, Berlin, Heidelberg.

Poostchi, H., Borzeshi, E.Z., Abdous, M., Piccardi, M. (2016). PersoNER: Persian Named-Entity Recognition. In Proceedings of COLING 2016, pages 3381–3389, Osaka, Japan.

Rasooli, M.S., Kouhestani, M., Moloodi A., (2013). Development of a Persian Syntactic Dependency Treebank, NAACL, pages 306-314, Atlanta, Georgia, USA.

Rouhizadeh, H., Shamsfard, M., Rouhizadeh M., (2019). Knowledge based word sense disambiguation with distributional semantic expansion. Widening NLP Workshop, ACL2019.

Sabeti, B., Hosseini, P., Ghassem-Sani G.R., Mirroshandel, S.A., (2016). LexiPers: An ontology based sentiment lexicon for Persian, In GCAI 2016. 2nd Global Conference on Artificial Intelligence, pages 329-339.

Shahshahani, M., Mohseni M., Shakery A., Faili H., (2019). PEYMA: A Tagged Corpus for Persian Named Entities, *Journal of Signal and Data Processing (JSDP)* 16 (1) :91-110.

Shamsfard, M. (2011) Challenges and open problems in Persian text processing. In: 5th Language & Technology Conference (LTC).

Shamsfard M., Hesabi A., Fadaei H., Mansoory N., Famian A., Bagherbeigi S., Fekri E., Monshizadeh M., and. Assi S. M (2010) Semi-automatic development of FarsNet; the Persian wordnet, In Proceedings of 5th global WordNet conference, Mumbai, India.

Shamsfard, M., Jafari, H.S. (2017) FarsVajeh: The Persian lexicon. Technical report, NLP Research Lab, Shahid Beheshti University, Tehran, Iran.

Taghizadeh N., Faili H., (2016). Automatic Wordnet Development for Low-resource Languages using Cross-Lingual WSD. *Journal of Artificial Intelligence Research* 56(56):61-87.

Proceedings, pages 296–299
Paris, UNESCO Headquarters, 5-6 December, 2019.

Designing for Language Revitalisation

Steven Bird
Northern Institute
Charles Darwin University
steven.bird@cdu.edu.au

Abstract

How do we design technologies, places, and activities that increase the survival prospects of a threatened language? The answer begins and ends with people, especially those who speak the languages in question. Too often, technologies for capturing languages do not offer an effective value proposition that would encourage large-scale participation by linguistic minorities in documenting their languages. A case in point is the Aikuma mobile app for crowdsourcing oral language documentation. I discuss Aikuma before reporting new designerly approaches in the areas of storytelling and language learning. These new approaches address the same challenge as before, only with better value propositions, while promising to deliver language documentation as a byproduct.

Keywords: oral cultures, storytelling, language learning

1. Introduction

How can people from dominant cultures encourage linguistic minorities to keep their languages strong? Language resources and technologies do not appear to be slowing the pace of language loss. Moreover, in our rush to preserve languages we may recapitulate the causes of language endangerment. Our preservation technologies enact *our* agendas. Our agency overrides local autonomy and self-determination. We enter marginalised communities with money, technology, a global language: an intoxicating blend that leaves local people in no doubt about where the real power lies.

In this paper I report on three design innovations that I have been exploring over the past decade in a bid to answer the opening question. The first innovation is technologisation, or: 'I will make you a thing' (Section 2.). The second works to extend the reach of multiculturalism: 'I will recognise your language' (Section 3.). The third seeks to leverage technology to transform the interactions between speakers of threatened and dominant languages in any place where they are thrust together; in effect we ask: 'How do I show respect and behave appropriately?' (Section 4.).

Figure 1: Recording Tembé (Cajueiro, Pará, Brazil)

2. Capturing Languages

It is not difficult to record large quantities of audio in minority communities (Figure 1). The challenge is to ensure that recordings are interpretable – to know what was said and what it meant – especially in the presence of ambient noise and audience participation. With 'careful respeaking', the source is repeated phrase by phrase in a quiet place (Woodbury, 2003). With 'oral translation', it is interpreted sentence by sentence into a language of wider communication. We bypass the 'transcription bottleneck' (Figure 2).

The Aikuma app provides a text-free interface to support respeaking and oral translation, shown in Figure 3(a) (Hanke and Bird, 2013; Bird et al., 2014). Users press and hold the left play button to hear the next segment of audio source. They can press it multiple times to hear the same segment over again. They press the right record button to respeak or translate. This process continues until the source has been fully processed. Aikuma generates a second audio file, time-aligned with the source. It supports playback of the source or translation or the two interleaved. The target language audio can be transcribed, resulting in audio, phrase-aligned to a written translation (Figure 3(b)).

The Aikuma app supports peer-to-peer file sharing, enabling us to demonstrate the concepts of storage and transmission to people who live far off-grid in remote villages and who have never experienced the Internet or digital archiving.

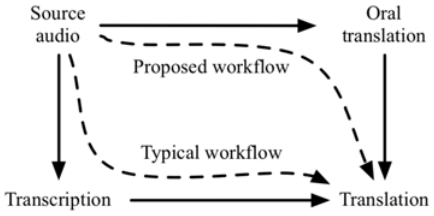
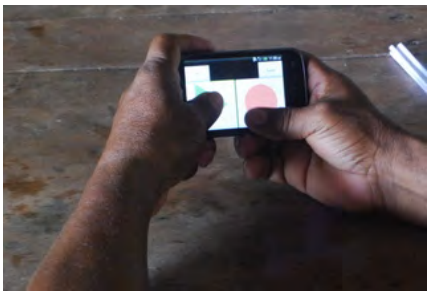
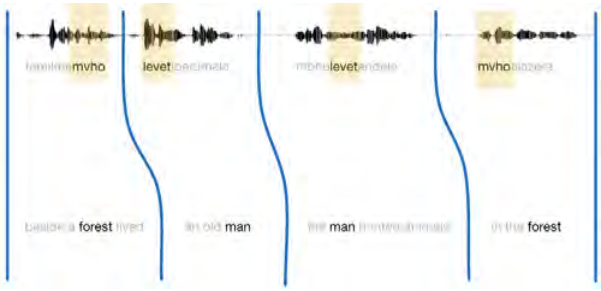


Figure 2: Bypassing the Transcription Bottleneck



(a) Translating into Portuguese



(b) Bilingual audio phrase-aligned with translation

Figure 3: Recording and translating using the Aikuma Android app

The Aikuma app has been extended with support for elicitation and editing (Blachon et al., 2016), and with support for images and gestures (Bettinson and Bird, 2018).

The premise of Aikuma was crowdsourcing, yet this did not happen. We had no value proposition for participation (Bird, 2018). Why should speakers of endangered languages use this app?

Aikuma was designed with two audiences in mind: elderly people who command the 'ancestral code', and unborn generations who want to revitalise their ancestral language. As such, it did not resonate with the very people who would need to animate the work in the present generation. In my experience in many minority communities, local people prioritise their economic prospects over their linguistic heritage. They are far more likely to ask for help with the dominant language than to ask for help with documenting a local language.

In our search for an effective value proposition, my colleagues and I tried to articulate a guiding vision. There were plenty of organisations promising to *halt* or to *slow* the tide of language loss, visions that seemed alternatively unrealistic or uninspiring. In time, we settled on a new vision, to *create a world that sustains its languages*. We inaugurated the Aikuma Project in 2015 with an initial task of designing a new storytelling format, returning to the meaning of Aikuma as 'meeting place'. At first we adopted the name 'treasure language', proposed by the Rama people of Nicaragua as a more positive alternative to terms such as threatened, disappearing, or dying language. In 2019, reflecting on the goals of celebrating and connecting, we changed the name to 'Language Party'.

3. Celebrating Languages

A *Language Party* is a community gathering where people celebrate locally-spoken languages and recognise those who are keeping them strong. Local storytellers share tales in their mother tongue then translate them into the dominant language of the audience (Figure 4).

Storytellers include indigenous people, migrants, expatriates, and refugees. They are not professional performers, but living conduits for languages that are little known. By speaking and sharing their languages, storytellers forge a deeper connection to their community, enriching its common life. Storytellers and story-listeners renew their resolve to keep their languages strong. Everyone comes to belong in a new way. Audience members have described the experience as 'an awakening', 'unexpectedly enjoyable and inspiring', 'moving and compelling' and 'a privilege to be part of'.

We have held Language Parties in places with strong linguistic diversity, starting in the San Francisco Bay Area, expanding across Australia, and spreading to many other places during the International Year of Indigenous Languages in 2019 (Figure 4(b)). Many events have been recorded, and the stories and their translations may in time come to be treated as a corpus in its own right.

The effectiveness of a Language Party lies in the storytellers, in the connections they make as they prepare for the performance, and in the chemistry that develops between storytellers and story-listeners.

Language parties extend the multiculturalism movement beyond cuisine and costume into a space that is a primal index of identity. We believe that many of the world's languages can be sustained if we adapt our urban places to embrace diversity, creating culturally safe spaces where people do not need to forget who they are in order to belong.

The approach works because it connects with the struggle of minority groups for *recognition* (McBride, 2013). Recognition Theory explains how personal identity is shaped by recognition, and how non-recognition inflicts harm on ethnic and linguistic minorities 'imprisoning someone in a false, distorted, and reduced mode of being' (Taylor, 1994, 25).

The framework of recognition also helps explain the popularity of technologies that are claimed to 'save languages' (Arnold, 2016), but whose power derives not from their content but from the recognition they bestow, in this case, recognition in the digital realm.

Both language apps and language parties leverage the prestige of the dominant culture in order to create symbolic spaces for the speakers of threatened languages, who are invited to enter and perform.

This is not to deny the significance of the approaches but to account for it, and to highlight the ongoing need to decolonise the intersection points of threatened and dominant languages. We conclude by discussing another space for design, beginning with the question: what happens when we take seriously local desires to participate in the mainstream economy and to learn the dominant language?

(a) Storytellers from Language Parties

(b) Language Parties: Prospective, Planned, and Confirmed

Figure 4: 'Stories in the original languages told by people who live in our midst' (`languageparty.org`)

4. Learning Languages

Threatened languages do not exist in isolation but in competition with a locally dominant variety. Speakers may use the local language at home and in the marketplace, switching to a dialect of the dominant language at school or in the workplace or when travelling to the provincial capital. The domains of use demonstrate the prestige indexed by each language (Fishman, 2001). When people representing the dominant culture enter this contested space, it is usually for a well-defined purpose and they generally speak the dominant language. However, there is another possibility, as we see in a remote community in Western Australia:

> The desire of non-Indigenous people (such as teachers, nurses and other community workers) to learn a Pilbara language was recognised as having the potential for positive flow-on effects throughout the community, in terms of improved provision of key services (especially in the health and education spheres), as well as increased awareness of Indigenous people's language rights. Both outcomes increase the prestige of Pilbara Aboriginal languages and create space within the broader community for language revitalisation to occur. (Dixon and Deak, 2010, p126)

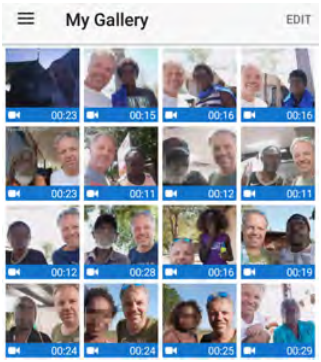
There is a risk of recolonisation when outsiders appropriate the local language. However, in places where there is already a long history of contact, a newcomer's efforts to learn language can be a welcome form of recognition. A choice phrase or greeting creates a 'moment of connection' (Galliford, 2010). Community engagement and language learning can take place concurrently (Christie, 2008). In learning the local language, outsiders remind themselves that they are in someone else's place, and acknowledge "the freedom of [local] people to lead the kind of lives they have reason to value" (Sen, 1999).

I have begun to explore this approach by learning Kunwinjku, an indigenous language of northern Australia spoken by 2,000 people. In the course of this work I have appropriated general-purpose mobile technologies to support my own oral language learning (Bird, 2019). The first is for learning the culturally appropriate way to address people, such as a kinship term, by capturing a selfie and recording a bilingual conversation about how we address each other (Figure 5(a)). The second is for efficiently capturing key vocabulary and phrases for achieving the task that justifies the outsider's presence (Figure 5(b)). The third is for obtaining 'comprehensible input', speech just beyond one's current level where one can leverage physical context to make meaning without access to translation (Krashen, 1981). These methods create bilingual resources which serve language learning *in either direction*. I used these methods to support my learning of Kunwinjku, and local people used them in learning English.

5. Conclusion

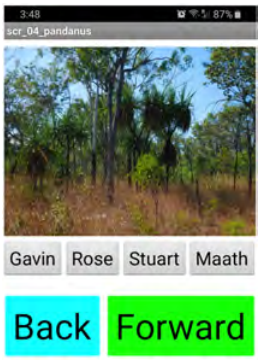
Most technology panaceas aimed at 'saving languages' are driven by hyperbolic valorisation, capturing language with minimal regard for local people and their struggles (Hill, 2002). There has been no theorisation about how language capture technologies reverse language shift, and no systematic evaluation of their effectiveness. There are other opportunities for language technologies: to support learning of threatened languages, and to support automatic processing of the low-prestige varieties of dominant languages that these people may already use to engage the outside world. Instead of treating the speakers of threatened languages as mere conduits, they could be viewed as collaborators or even commissioners of language work, as has occasionally been advocated by linguists (Rice, 2011; Sapién, 2018). This is consistent with their entitlement to autonomy and self-determination as set out in the United Nations Declaration on the Rights of Indigenous Peoples.

How then can people from dominant cultures encourage linguistic minorities to keep their languages strong? We must engage with local people in addressing the causes of language shift, and in strategising about which domains can be reclaimed from the dominant language (Fishman, 2001). I have outlined two highly generative responses to the question. The format of 'Language Parties' brings speakers of stigmatised varieties into places of high culture where recognition leads to pride and a new sense of belonging. Learning of a stigmatised language by outsiders, with permission, demonstrates deep respect for local knowledge authorities. Language resources – and even language vitality – may emerge, but we do not lose sight of the speech community and their sovereignty.



(a) Learning terms of address by capturing selfies and recording a short bilingual dialogue

(b) Learning key phrases by capturing spoken terms and translations



(c) Learning linked to descriptions of places

Figure 5: Appropriating Simple Apps to Support Oral Language Learning

6. Acknowledgements

I am grateful for support from the US National Science Foundation, the Australian Research Council, the Firebird Foundation, the Aesop Foundation, Warddeken Land Management Limited, and the Australian Indigenous Languages and Arts Program. I am indebted to many collaborators including Florian Hanke, Oliver Adams, Haejoong Lee, Lauren Gawne, Katie Gelbart, Isaac McAlister, Sangyeop Lee, Mat Bettinson, Alexandra Marley, Dean Yibarbuk, Laurent Besacier, David Chiang, Manuel Maqueda, Robyn Perry, Nadia Chaney, and Jennifer Pinkerton.

References

Arnold, C. (2016). Can an app save an ancient language? *Scientific American*.

Bettinson, M. and Bird, S. (2018). Image-Gesture-Voice: a web component for eliciting speech. In *Third Workshop on Collaboration and Computing for Under-Resourced Languages*.

Bird, S., Gawne, L., Gelbart, K., and McAlister, I. (2014). Collecting bilingual audio in remote indigenous communities. In *Proceedings of the 25th International Conference on Computational Linguistics*.

Bird, S. (2018). Designing mobile applications for endangered languages. In *Oxford Handbook of Endangered Languages*. Oxford University Press.

Bird, S. (2019). Designing for participation: mobile-assisted oral language learning in an Australian Aboriginal community. Manuscript.

Blachon, D., Gauthier, E., Besacier, L., Kouaratab, G.-N., Adda-Decker, M., and Rialland, A. (2016). Parallel speech collection for under-resourced language studies using the LIG-Aikuma mobile device app. In *Proceedings of the Fifth Workshop on Spoken Language Technologies for Under-resourced languages*, pages 61–66. Elsevier.

Christie, M. (2008). Yolngu studies: A case study of Aboriginal community engagement. *Gateways: International Journal of Community Research and Engagement*, 1:31–47.

Dixon, S. and Deak, E. (2010). Language centre as language revitalisation strategy: a case study from the Pilbara. In John Hobson, editor, *Re-awakening Languages: Theory and Practice in the Revitalisation of Australia's Indigenous Languages*, pages 119–30. Sydney University Press.

Fishman, J. A. (2001). Why is it so hard to save a threatened language? In Joshua A. Fishman, editor, *Can Threatened Languages be Saved?: Reversing Language Shift, Revisited: a 21st Century Perspective*, pages 1–22. Multilingual Matters.

Galliford, M. (2010). Touring 'country', sharing 'home': Aboriginal tourism, Australian tourists and the possibilities for cultural transversality. *Tourist Studies*, 10:227–44.

Hanke, F. and Bird, S. (2013). Large-scale text collection for unwritten languages. In *Proceedings of the 6th International Joint Conference on Natural Language Processing*, pages 1134–38.

Hill, J. (2002). "Expert Rhetorics" in advocacy for endangered languages: Who is listening, and what do they hear? *Journal of Linguistic Anthropology*, 12:119–33.

Krashen, S. D. (1981). The "fundamental pedagogical principle" in second language teaching. *Studia Linguistica*, 35:50–70.

McBride, C. (2013). *Recognition*. Wiley.

Rice, K. (2011). Documentary linguistics and community relations. *Language Documentation and Conservation*, 5:187–207.

Sapién, R.-M. (2018). Design and implementation of collaborative language documentation projects. In *Oxford Handbook of Endangered Languages*. Oxford University Press.

Sen, A. (1999). *Development as Freedom*. Oxford University Press.

Taylor, C. (1994). *Multiculturalism: Examining the Politics of Recognition*. Princeton University Press.

Woodbury, A. C. (2003). Defining documentary linguistics. *Language Documentation and Description*, 1:35–51.

Proceedings, pages 300–303
Paris, UNESCO Headquarters, 5-6 December, 2019.

Contribution to the Universal Dependencies Treebank of Non-Standard Romanian Texts

Victoria Bobicev[1], Catalina Maranduc[2], Tudor Bumbu[3], Ludmila Malahov[3], Alexandru Colesnicov[3], Svetlana Cojocaru[3]

[1]Technical University of Moldova, [2]"Al. I. Cuza" University, [3]Institute of Mathematics and Computer Science of the Academy of Sciences of Moldova
[1]Chișinău, Moldova, [2]Iași, Romania, [3]Chișinău, Moldova

victoria.bobicev@ia.utm.md, catalinamaranduc@gmail.com, bumbutudor10@gmail.com, lmalahov@gmail.com, acolesnicov@gmx.com, svetlana.cojocaru@math.md

Abstract

Cultural heritage preservation is a non-transferable duty of any ethnic or social entity; it is the essence that defines each one of them among others. In our specific case of culturally significant literary works preservation, the task includes not only the digitization of old books to prevent their loss, but also the optical character recognition, transliteration of Cyrillic Roman texts and their annotation. We report our recent contribution to the development of the Universal Dependency Treebank (UD) which contains old texts, regional folklore and other non-standard texts from Moldova and Romania in Non-Standard Romanian UD corpus.

Keywords: Cultural heritage preservation, Romanian Cyrillic texts, digitizing old books, optical character recognition, transliteration, morpho-syntactic text annotation.

Rezumat

Păstrarea patrimoniului cultural este datoria netransmisibilă a oricărei entități etnice sau sociale, deoarece este esența care o definește și identifică. În cazul specific al conservării operelor literare semnificative din punct de vedere cultural, această sarcină include nu numai digitalizarea cărților vechi pentru a preveni pierderea lor, dar și recunoașterea optică a caracterelor, transliterarea textelor romane chirilice și adnotarea lor. Raportăm contribuția noastră recentă la dezvoltarea Treebank-ului de dependențe universale (UD) care conține texte vechi, folclor regional și alte texte non-standard din Moldova și România.

1. Introduction

Digitisation, preservation and online access to historic literary and cultural treasures are listed among the priorities of the Digital Agenda for Europe. The actions undertaken by the EU include the development of the European Digital Library *Europeana*[1], supported by the EU Program for Culture. Multiple European research groups and laboratories addressed various problems of creation of linguistic resources by digitisation and recognition of historic and literary heritage (Moruz et al., 2012) through different European projects. Unfortunately, the scientific centres of the Republic of Moldova aren't involved in these actions in spite of their efforts in this domain.

The Government of the Republic of Moldova approved the National Strategy for the development of information society "Digital Moldova 2020" and the Plan of Actions for implementation of this Strategy: the Program "Creation, development and evaluation of the digital content in the RM in 2016 - 2020".

The main aims of the cultural policy for the spaces where the Romanian language is spoken include the study, digitization and preservation of its heritage. The digitization process requires solving a series of problems related to the recognition, editing, transliteration, interpretation and reception of printed Romanian texts in both Latin and Cyrillic alphabets.

Working on these tasks for the Romanian historical linguistic heritage means solving a number of specific problems, namely: the large number of periods in the evolution of the language, the small volume of resources widely distributed, the great diversity of alphabets used in their printing, in particular, mixed Cyrillic-Latin "transition alphabets", the lack of tools for the correct recognition of Cyrillic letters from different historical periods, as well as the lack of lexicons suitable for the period of printing of the resource.

In order to overcome the abovementioned problems a platform has been created which integrates a set of software components for image processing, text recognition and transliteration into modern Latin spelling. It has been adapted for the recognition and transliteration of texts from different historical periods, and for the differences in evolution of the alphabets used for Romanian language printings in Romania and in the present territory of the Republic of Moldova.

2. Our Heritage

We work with old Romanian books in the Cyrillic script.

[1] https://ec.europa.eu/digital-single-market/en/europeana-european-digital-library-all

The researchers of Romania and of The Republic of Moldova have the same problem. The two countries constituted a single state in the past, the historical documents (written in old Romanian Cyrillic) are common, and the regional variants of Romanian spoken in the two countries, with minor differences, are mutually understandable.

The starting point of our work is the scanned text, i.e., the text presented in the form of page images. The sources of these text images are the electronic libraries of texts in this form, e.g., Bucharest Digital Library[2], National Library of Moldova[3].

Table 1 lists the linguistic sources we have been working on recently. The sources are of different historical periods starting with the oldest ones printed in the very first printing houses situated in Moldova in the XVII-XVIII centuries.

Romanian Cyrillic fonts, especially of the selected epoch, are much less variable than Latin ones. The usage of the Cyrillic script is connected with the Slavonic liturgical language of the Orthodox Church.

XVII century	Noul Testament, 1648
XVIII century	Fiziognomie, 1785 Ducere către aritmetica, 1785 De obste Gheografie, 1795 Așezamânt, 1786
XIX century	Epistolariu, 1841 Gramatica românească, 1835 Legiuirea Caragea, 1818
XX century	Folclor din părțile codrilor, 1973 Colecții de reviste 1950-1992

Table 1: Scanned, recognized and transliterated books.

Figure 1 presents small fragments of scanned text of different periods.

3. OCR: problems and solutions

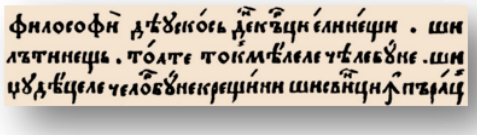
Post-processing of digitized text is a complex task. To solve it, we are developing software that supports expert's efforts in improvement and analysis of the recognized texts. The highest priority task of post-processing is to minimize errors in the recognized text.

The conversion of historical documents from the paper to accessible and searchable electronic form meets two obstacles that are not fully cleared till now. Nowadays state-of-the-art in OCR guarantees relatively good results only on modern texts. For historical typography, results are worse for several causes. Historical fonts vary even in one book, and are less readable. Old paper introduces speckles and distortions. Linguistic components and resources of modern systems don't often know the peculiarities of historical language variations. Each text yields its own specific mix of features and problems, which implies that the quality of OCR for historical documents may vary from perfect to almost unacceptable. The second general problem is produced by the historical orthography and language changes. Most users of digital libraries don't have a good command of old language and desire to use the modern orthography at their search. Any word can have numerous variants in the historical documents because of language evolution and lack of orthography standardization. To get satisfactory replies at search, it is necessary to bridge the gap between modern and old orthography. Availability of texts in original historical orthography differs considerably for different scripts. For example, Romanian Cyrillic script of the 18th century has glyphs that are not supported by most OCR programs.dsf

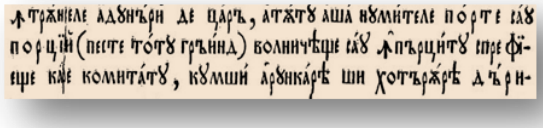
Technology for recognition of the historic and linguistic Romanian heritage printed in the Cyrillic script in the 17th–20th centuries is supported by a pack of the following tools and utilities:

- Alphabets for ABBYY FineReader (AFR).
- Dictionaries (word lists) for AFR.
- Recognition patterns as trained under AFR.
- Selection utility to start AFR with the alphabet, dictionary, and templates corresponding to a specific epoch and location.
- Virtual keyboard.

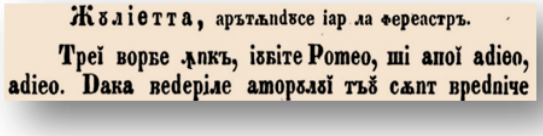
XVII century



XVIII century



XIX century, mixed alphabet



XX century

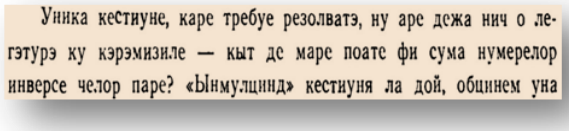


Figure 1: The small fragments of the texts from different centuries.

[2] http://www.bibnat.ro/
[3] http://www.bnrm.md/

The recognition of texts of the 18th-19th century resulted in WER (Word Error Rate) of 3−4.5% and the WER of the 17th century is more than 6%. Figure 2 presents a fragment of Cyrillic text from XVII century after its recognition. In spite of the fact that all letters a clearly seen it is still barely readable by a modern Romanian speaking person due to the specific alphabet which mix Latin, Cyrillic and some Greek-looking letters.

To solve the problem of multiple character sets for the old texts we developed historical alphabets and sets of glyphs recognition templates specific for each epoch. The dictionaries in proper alphabets and orthographies were created in order to minimize the error rate. In addition, virtual keyboards, fonts, transliteration utilities, and other tools were developed for the researchers of old documents.

A special interface is created for the selection of the historical period and the geographical region, where the text was printed. User can choose one of the following variants: Iasi, Bucharest, Târgoviște, Bălgrad (Alba Iulia), Uniev (Cernăuți), Sas Sebeș, Snagov or Buzău. Within a region the typography should be selected. For example, for Bucharest the system is trained in recognizing the fonts from the Royal Typography and that of the Bucharest Metropolitan Chair.

4. Romanization of Cyrillic

Once the scanned image was processed and the editable and intelligible Cyrillic text was obtained, the trans-literation process takes place.

Unusual fonts are difficult for perception even for professionals in linguistics. Therefore, solving the problem of textual cultural heritage dissemination supposes the development of tools for transliteration in common modern Romanian alphabet.

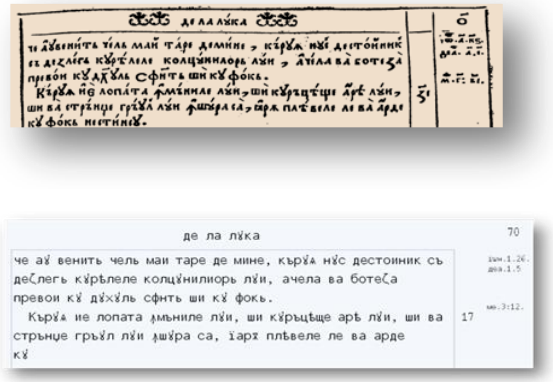


Figure 2: A fragment of a scanned page from New Testament (1648) before and after OCR.

The first problem is presentation of recognized Cyrillic text in computer, especially for transitional (TR) and Romanian Cyrillic (RC). In fact, only three fonts in the whole world have old Romanian Cyrillic letters: Kliment STD, Unifont, and Everson Mono only since 2009.

Simplified Romanian Cyrillic script (SRC) of the mid 18th century till 1830 is characterized by two substantial differences from that of the older time. Each period has its specifics and needs specific processing.

Transitional alphabets were used in the Romanian typography since 1830 and until 1860-1870 (Cazimir, 2006). They can be characterized by regular many-to-one mapping of Romanian Cyrillic letters to the mix of Latin and Cyrillic letters. This mapping could be expanded further to modern Latin Romanian script; slightly different orthography poses an obstacle. The existence of such mapping distinguishes the Romanian Cyrillic and transitional scripts from Moldavian Cyrillic script (MC) that cannot be regularly mapped to the modern Latin script (Ciubotaru et al., 2015). The solution of these problems for the Republic of Moldova faces specific difficulties: the existing resources are scarce and they were printed in diverse alphabets.

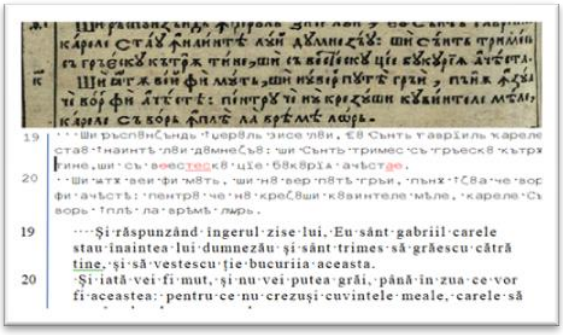


Figure 3: Three forms of an old document: scanned picture, recognized editable Romanian Cyrillic text and the text transcribed in Latin letters.

The transliteration of the Moldavian Cyrillic to Modern Romanian Latin (MRL) was discussed in details in (Boian et al., 2014). The method is rule based; three groups of rules were created manually. Most letters (26 of 31) can be mapped one-to-one as, for example, **ш** to **ș**; **ц** to **ț**. Three letters (**г**, **к**, **ч**) can be transformed using context-dependent rules. The letter **ы** may be transformed in either **î** or **â** in accordance with the rule of the Romanian language. The letter **я** is the most difficult case that can't be fully solved without access to dictionaries. Rules are mostly heuristic and statistical, and more than 20 rules do not cover all cases. This situation exists because MC was not thoroughly designed but is an *ad hoc* mapping of Romanian sounds to the Russian letters.

The transliteration algorithm of Romanian Cyrillic and transitional alphabets contains mostly simple rules; the letters in these alphabets are less ambiguous. The Romanian Cyrillic script reflected the word composition the most accurate. The accuracy of conversion is up to 95% for Moldavian Cyrillic, up to 96% for Transition alphabets and up to 98% for Romanian Cyrillic.

5. Annotation and conversion in UD

The next step after the texts were transliterated was their enrichment with the linguistic information. The

texts were automatically processed at UAIC[4] by the Robin-hybrid POS-tagger (Simionescu, 2011), using MULTEXT East project PoS tags (Erjavec, 2004). The set of morpho-syntactic tags for Romanian language developed during the MULTEXT project consisted of 614 tags. This set was quite large with the detailed description of the specifics of Romanian morphology. We simplified the tags keeping 450 tags from the main set and adding around 100 tags to annotate specific elements of the old language. Table 1 contains an example of text enriched by the morpho-syntactic tags.

Old Romanian **Cyrillic**	**Modern Romanian**	**Morpho-syntactic tag**
Їwсиф	Iosif	Npmsrn
фѳци	fugi	Vmis3s
кѳ	cu	Spsa
ЇС	Iisus	Npmsrn

Table 2: A fragment of text from Noul Testament written in old Cyrillic transliterated in modern Romanian and enriched with morphological tags.

The accuracy of automate morphological tagging was 95 - 96% on various modern Romanian texts; on old texts it was considerably lower. We enriched the dictionary with old Romanian words and manually corrected the annotated old texts and by the bootstrapping method we created the non-standard gold annotated corpus.

Syntactic annotation was obtained by automate annotation by MaltParser and subsequent manual verification and correction by linguists. The convention of annotation is FDG (Functional Dependency Grammar), with labels of classical syntax, with numerous semantic sub-classifications of modifiers. The first texts of our corpus were annotated using the parser trained on UAICRoDepTb (UAIC Romanian Dependency Treebank) and the automate annotation had to be verified and corrected manually as its accuracy on the non-standard old texts was quite poor.

We annotate using dependency grammar formalism developed at UAIC which can be transformed in two formats: the modern syntactic system of Universal Dependencies (UD) with loss of semantic information and into a semantic annotation system by adding information.

The corpus annotated in the initial formalism is registered as UAIC-RoDia DepTB[5] (Romanian Diacronic Dependency TreeBank) and in the UD format is uploaded as a part of UD project Romanian-Nonstandard corpus[6].

The annotated part of the corpus is growing rapidly and has now 15843 sentences and 318869 tokens containing old texts, (1592-1818), and folklore from Romania and Republic of Moldova. The accuracy of automate annotation of the text in this treebank is around 80% and we are working in order to obtain the accuracy over 90% by increasing the gold annotated corpus verified and corrected manually.

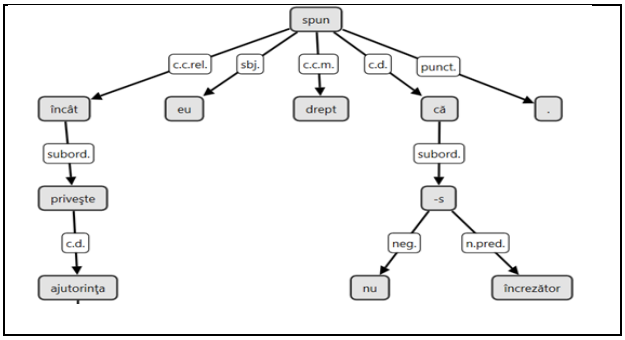


Figure 4: A fragment of a sentence with syntactic annotation opened in a graphic editor for the annotation correction.

6. Conclusion

The paper describes our contribution to the development and enrichment of an annotated Romanian corpus which contains old texts, regional folklore and other non-standard texts from Moldova and Romania. Digitization of old documents poses interesting scientific challenges; however NLP for such texts tends to be carried out in isolated and sparse research groups, and the resulting products are often in different formats and standards. We are working on digitizing old Romanian texts improving their optical character recognition, transliteration in modern Romanian script, their annotation to finally include them in UD repository in a common standard.

7. Bibliographical References

Boian, E.; Ciubotaru, C.; Cojocaru, S.; Colesnicov, A; Malahov, L. (2014). Digitizarea, recunoaşterea si conservarea patrimoniului cultural-istoric. Akademos, Nr. 1(32), 2014, pp. 61–68.

S.Cazimir (2006). The transitional alphabet. - Bucharest: Humanitas, ISSN 973-50-1401-7. (In Romanian)

C.Ciubotaru, S.Cojocaru, A.Colesnicov, V.Demidov, L.Malahov (2015). Regeneration of Cultural Heritage: Problems Related to Moldavian Cyrillic Alphabet. International Conference "Linguistic Resources and Tools for Processing the Romanian Language". Eds: D.Gîfu, D.Trandabăţ, D.Cristea, D.Tufiş. pp. 177-184.

Erjavec, T. (2004). MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora.In Proc. of the Fourth Intl. Conf. On Language Resources and Evaluation, LREC'2004, ELRA http://nl.ijs.si/ME/Vault/CD/docs/mte-d11f/

Moruz, M.; Iftene, A.; Moruz, A.; Cristea, D. (2012). Semi-automatic alignment of old Romanian words using lexicons. In: Proceedings of the 8th International Conference „Linguistic resources and tools for processing of the Romanian language", Iaşi, Editura Universităţii „A.I. Cuza", p. 119–125.

Simionescu, R. (2011). Hybrid POS Tagger. In: Proceedings of "Language Resources and Tools with Industrial Applications", Workshop Eurolan 2011 summer school.

[4] Alexandru Ioan Cuza University, Iaşi, Romania

[5] ISLRN 156-635-615-024-0

[6] https://universaldependencies.org

Proceedings, pages 304–307
Paris, UNESCO Headquarters, 5-6 December, 2019.

Translation Commons: No Language and No Linguist Left Behind

Jeannette Stewart
Translation Commons
Cupertino, USA
jeannette@translationcommons.org

Abstract

This paper describes work conducted by Translation Commons to provide a reliable method of digitally rendering a language. It seeks to solve the problem that unrepresented languages have participating in a global communications network. We present a methodology that can easily be utilized by non-technical users to meet the particular needs of their language. Evidence of the success of this approach is cited with examples. The importance of this initiative is illustrated by recommendations for on-going and new projects.

Keywords: Translation Commons, Language Technology Project, Zero to Digital, indigenous communities, language status, language code, Unicode, Common Locale Data Repository (CLDR), University Outreach, Social Media, Global Language Digitization Initiative (GLDI), fonts, keyboards, digitization.

Περίληψη

Αυτό το άρθρο περιγράφει την εργασία που διεξήγαγε η μη κερδοσκοπική οργάνωση Translation Commons για να παρέχει μια αξιόπιστη μέθοδο ψηφιακής απόδοσης μιας γλώσσας. Επιδιώκει να λύσει το πρόβλημα ώστε οι μη αντιπροσωπευόμενες γλώσσες να μπορούν να συμμετέχουν σε ένα παγκόσμιο δίκτυο επικοινωνιών. Παρουσιάζουμε μια μεθοδολογία που μπορεί εύκολα να χρησιμοποιηθεί από μη τεχνικούς χρήστες για την κάλυψη των ιδιαίτερων αναγκών της γλώσσας τους. Η απόδειξη της επιτυχίας αυτής της προσέγγισης παρατίθεται με παραδείγματα. Η σημασία αυτής της πρωτοβουλίας αποδεικνύεται με συστάσεις για τρέχοντα και νέα έργα.

1. Introduction

Are all languages equal? Our language environment is teeming with life. An ecology of language involves devoted care and attention to conserve, guarantee sustainability, assure health and well-being with inarguable equality. Development of our digital domain has been unequal, endangering under-resourced languages. Strenuous effort is needed to future-proofing against exclusion and even extinction.

2. Translation Commons

Translation Commons is a nonprofit organization with a vision to help every language digitize to share equally in the benefits of a connected digital world, ensuring that **"no language and no linguist is left behind".** Our online platform provides free access to language tools, language training and sharing of educational knowledge for all languages. During 2019, we are a Social-Civil society partner to International Year of Indigenous Languages and have been actively working on three projects: a language technology project, a university outreach project and a social media campaign.

2.1 Language Technology Project

The language technology team agreed on the following objectives:

- **Enable** indigenous communities to use internet technology
- **Engage speakers** and especially young learners to appreciate and honor their culture
- Identify **technology for educating**
- Encourage speakers to **create content** and community with digital tools and standards

All in their own language.

2.2 Zero to Digital Document

Their first project was a step-by-step document as technical guidelines on how to enable a language to be online. The document Zero to Digital is the brainchild of 4 of the most committed volunteers in the digitization of languages, all experts in their respective fields: Craig Cornelius, Craig Cummings, Deborah Anderson and Lee Collins. You can download it at translationcommons.org.

2.3 Objectives of the Document

The guidelines document helps communities:

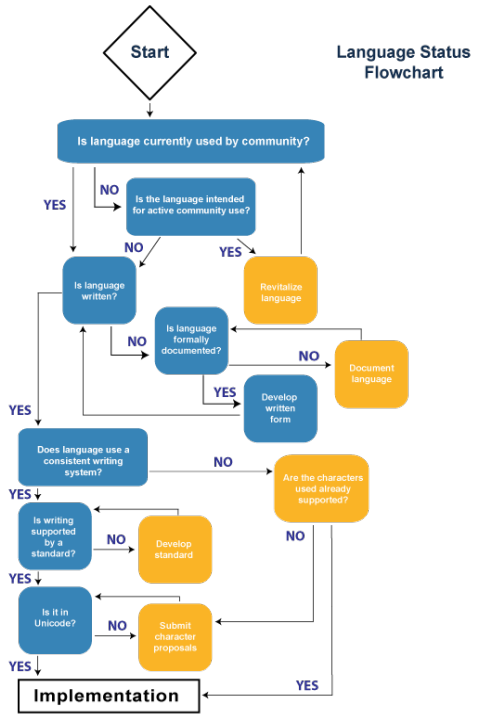
- **Evaluate** their language's situation with respect to digital support
- **Suggest approaches** to develop basic digital language support
- **Encourage practical usage** of language tools, even without formal language documentation, grammar, and educational standards
- **Point indigenous communities to available tools** and techniques to build digital capabilities.

- **Engage community members** in decisions and process
- **Connect communities to standards** and technology professionals

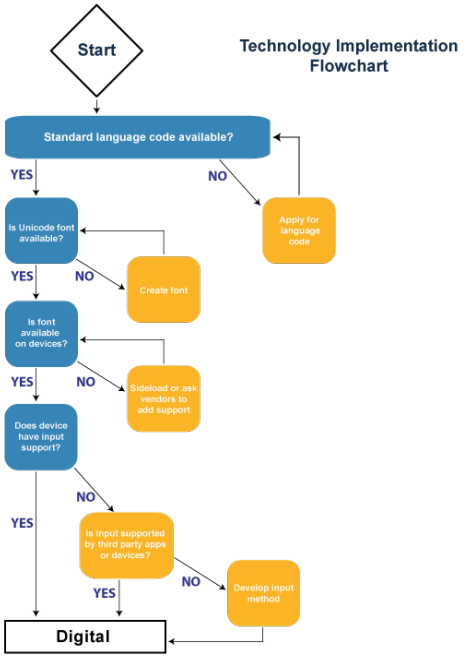
2.4 Zero to Digital Flowcharts

With two main flowcharts to show the basic workflows in digitization, it is very easy to follow and has a more technical section at the end for the technical implementation. By using questions answered with Yes-No, the communities are guided to the next step they need to undertake with the available resources and people. It offers concrete actions for each response, a guide to perform specific tasks, pointers to resources and tools and suggestions for next steps.

1.4.1 Language Status Flowchart
This workflow describes the steps to determine the current status of a language in preparation for using it online. The aim is to determine the current level of support for your language on computers and mobile devices. It includes suggested next steps to help you take your language online. You do not need a technical background to answer these questions.



1.4.2 Technology Implementation Flowchart
When you have determined the status of your language, use this workflow to get started with the technology available to take your language online.



Any written language used in digital devices must therefore include:

- a standard encoding system for characters, for example Unicode,
- rendering systems plus fonts for characters used in the language, and
- methods or applications to render the code points to desired media.

1.4.3 Unicode
Unicode standardization covers many of the text processing aspects of a writing system and language. With font and input support, many things will just work. This includes most aspects of basic word processing, spreadsheets, and email.

1.4.4 Common Locale Data Repository (CLDR)
The Common Locale Data Repository (CLDR) provides "key building blocks for software to support the world's languages" by collecting useful information for different locales (language and country). This data can provide the names of languages, countries, months, weekdays, and other information. It also enables locale-aware formatting of date, time, numbers, and other commonly formatted information. Although CLDR data is not required for basic text communication in indigenous languages, this information enhances language functionality. Almost all tools such as email, texting, social media, and so on will work well when fonts and keyboard are present.

This information is used by programmers to create output for online applications in specific context such as localized calendars, spreadsheets, numeric output, menu selections, and other user interface contexts.

3. University Outreach Project

For our second project, we reached out to the Language Departments of universities around the world and asked them to create an awareness event in their institution. To host classroom presentations, lectures, workshops and conferences during Fall 2019 to discuss cultural diversity, focusing on indigenous languages. We created a multicultural intern group and reached out to over 650 universities, of which many created events. They are still creating events until the end of the year and we just met with a group of them to brainstorm on next steps for the years to come.

4. Social Media Campaign

Our third project of a social media campaign reached 200,000 people with over 15% engagement. This was done by volunteers with no budget. The significance of this success has made us now aim for 1M viewers.

4.1 Campaign objectives

The campaign was named "Sharing Inspiration":

- Helping indigenous groups to inspire each other
    - Language education
    - Leveraging modern technology
    - Cultural pride and resilience
- Inspiring the audience to see the potential of indigenous speakers
    - Entrepreneurship and innovation
    - Indigenous knowledge (sciences, historical knowledge)
    - Successful language revitalization projects

5. Success Stories

5.1 Cherokee Nation

The Cherokee Nation in North America. From creating their syllabary in 1836 to contracting out the creation of their fonts they were always committed to the latest technology, from typewriter balls to early mobile and today they are fully digitized, from operating systems and desktop applications to texting, mobile apps like maps and mango, social media and everything else we do in the main 20 languages.

5.2 ADLaM

Another example is from the early 1990s when two teenagers created a writing system for the African language of the Fulani people. Their work became "Bindi Pulaar" and eventually "Adlam". An article entitled "The Alphabet that Will Save a People from Disappearing" describes how Adlam became widespread across many African countries. This had a huge impact on a couple of technologists who helped the brothers digitize it. The ADLaM journey started first with handwriting, then implemented it as a font encoding with Arabic code points. Technical challenges and support by technology vendors are enabling rapid growth of this alphabet in a previously under-served language community implementation. The challenges are many as the newly standardized script meets the technical infrastructure of the Internet. ADLaM holds great potential to improve literacy for millions as it is spreading rapidly in more than 20 countries.

6. What's next ?

Digitization is practically possible when the community is fully committed.

6.1 Internet Connectivity

There are over 4,000 written languages (many more oral languages) in the world today and only 1,000 of them are supported online. This is because most of these languages are not yet digitized.

What would the world be like if we were all online? With so many international agencies and companies working towards 100% connectivity, we ask: In what language will the next 7 billion people use the internet? This is an answer only we can give, all of us in this room and many more from the communities themselves, the academic world, the language professionals and most importantly our language technologists.

Bringing connectivity to emerging countries will solve many problems and create great opportunities. BUT connectivity alone will not allow people to be online unless their language is digitally supported. To fully partake in the benefits of online connectivity, people will need to be able to type, speak and read the online material in their own language. What if we could bring together everything we all learned and work at a global scale?

7. Global Language Digitization Initiative

Translation Commons' mission is to take all necessary steps to bring every language online so emerging countries can benefit from connectivity.
It is a vision we all share and we have all worked towards. But we are each an expert in a different field and it is only when we combine our knowledge, passion and fortitude, we will accomplish our vision.
Let's have our technologists create the technical support and our academics the linguistic support. Let's have our educators and researchers work on teaching the next generations and let's get the language professionals to embrace and include new linguists from thousands of languages.
We have created the **Global Language Digitization Initiative (GLDI)**. Our network of volunteers comes from all language specializations, our educational network of universities and schools is rapidly expanding and our platform offers free training and practice for new linguists, putting Translation Commons in a unique position to be the Language Technology partner among the team of organizations striving for 100% global connectivity.

7.1 'It takes a village to raise a child'

Inspired by this famous African proverb, we acknowledge the critical importance of nurturing newly digitized languages, parenting them, as it were, allowing them to grow in safety, educating them and providing for their needs until they reach sufficient maturity to take care of themselves.

Our curiosity about these children's future leads us to ask:
"What if we could make a bigger difference than anything we've done so far?"
"What if we could honor languages of communities past and present?"
"What is we could work on something so important that it could lead us to solve some of the world's toughest problems?"

We can do this by partnering, creating coalitions and working together: "a broad, multi-stakeholder alliance for digital cooperation" as suggested by the UN report on June 2019. This is the best way to achieve the UN Sustainable Development Goals.
Today Translation Commons is two thousand volunteers strong, a million tomorrow!
Everyone in Translation Commons is committed to this vision and we pledge to be the coordinating force between all of us, to connect and channel our expertise where, when and how it is needed.

7.2 GLDI Phases

For a language to be supported on the internet, a series of steps must occur:
Language Technology

- Create the stack of documents on how to digitize. Language Digitization means encoding their script so it can be recognized digitally; creating software for their keyboard and other means of input; creating templates for them to start compiling digital archives of voice and text; creating digital dictionaries and digital courses for children to learn their language through modern tools.
    - Zero to Digital, a step-by-step guide to implement language digitization
- Reach out to communities and assess their digital needs
- Guide the communities through implementation of language digitization.
-

Linguistic and Cultural Support

- Create teams of Language Ambassadors to connect communities with local educational institutions and schools
- Create groups of Project Managers to liaise between communities and various expert organizations that will help them digitize
- Provide education on internet safety in their own language
- Create collaboration between universities to share knowledge and research
    - Lecture exchange
    - Curriculum enhancement
    - Incorporate new digitized languages in standard settings

Create Economic Equal Opportunities
Facilitate the creation of new jobs by training translators to help communities share knowledge and information through translating from and into their language

- Offer access to free tools for linguists already on TC Platform
    - Computer-assisted Translation
    - Translation Memory technology
    - Machine Translation
    - Glossaries/Terminology Management
- Offer free tutorials, courses, videos for linguists through the TC Learning Center
- Create a marketplace for linguists of new languages for
    - Internships and jobs with industry
    - Internships and jobs with language companies

8. Conclusion

The Zero to Digital document presents the global language community with a broad vision to bring together many diverse organizations, communities and institutions. Progress can only be achieved when all of the aforementioned are bridged by a coordinating body. Translation Commons is pledged to build and maintain this bridge, providing the means for all language users to share equally in the opportunities offered by digital life.

9. Copyright

Proceedings, pages 308–310
Paris, UNESCO Headquarters, 5-6 December, 2019.

Wa7 szum'in'stum' ti nqweleutenlhkalha
Technology Help and Hindrance in Indigenous Language Revitalization

Lorna Wanosts'a7 Williams
Lil'wat, Professor Emerita University of Victoria, First Peoples Cultural Foundation
Victoria, BC Canada

Abstract

We hold our languages close to our hearts, they are precious to us, that is the closest translation of the words expressed in my language in the title. This article will give some illustrations of how technology helps and also hinders keeping Indigenous languages thriving. It is an opportunity to keep every language born of the land living and thriving.

Wa7 szum'inum' ti nqweleutenlhkalha. Huz kan qwelutentumulh eltsa kwa maysneml kwa nukw'uns ti nqwelutenlhkalha ti teknoaatsihas. Wa7 ka nukw'cala, wa7 muta xelilcs t'aqmens. Tuxlhoamas ka nuk'anems ti nqwelutens ti tmicwkalha.

This is the story of people, ucwalmicw, revitalizing their languages and the opportunities and challenges faced with technologies. Indigenous peoples have always used the technologies as they developed in the world. For example people in the north used radio to connect over long distances or within their communities to tell their stories and share news like they did in their community circles. People who thought that they no longer had songs from their ancestors found samples of their songs and stories on wax cylinders in far off places and brought them home.

2019 was designated the United Nations Year of Indigenous Languages, there are few days left, the designation has helped with focusing attention on Indigenous Languages around the world. In Canada there is finally a Federal Indigenous Languages Act intended to support the reclamation, revitalization, maintaining and strengthening Indigenous languages in Canada. The Act received Royal Assent on June 21st, 2019. Both levels of governments made a commitment to respond to both the United Nations Declaration on the Rights of Indigenous peoples (UNDRIP) and the Truth and Reconciliation Commission (TRC) recommendations to address the revitalization and reclamation of Indigenous Languages. In B.C. the UNDRIP bill was tabled in the legislature on 24 October 2019 and passed in November 2019. Carrying out this new mandate requires intense and extensive planning at every level; as every institution in the current government structure has had a part in the demise of Indigenous languages. It means developing new policies; revising and amending existing policies; establishing functioning spaces in all institutions for Indigenous languages, and shifting ingrained habits of thinking, beliefs and practices regarding Indigenous languages and Indigenous knowledge. The task in 2019 and beyond is to reverse the shift from a state of decline to one of vitality and active language use.

As residents of these lands, where all Indigenous languages were born, the challenge is to turn and face the effects of colonial history honestly, with courage and resolve to take the path towards renewal. Canada is not alone in tackling language and culture shift. Language annihilation has been a key to taking over land and asserting power over Indigenous peoples the world over. Indigenous languages have been under threat from the beginning of the arrival of explorers, fur traders, gold miners, Christian missionaries and settlers. The most powerful force in the eradication of Indigenous languages was education, especially the Residential schools, but the Federal Indian Day Schools had the same policies of not permitting Indigenous languages to be spoken. Until very recently there has been no support to keep the Indigenous language thriving. Those that continue do so are due to the efforts of the Indigenous peoples of the language communities and the assistance of some allies in the linguistics community. See First Peoples' Cultural Council (FPCC) website for the Status of Indigenous Languages reports (*Report on the Status of B.C. First Nations Languages 2018)*. Language is power—it carries the histories, knowledge and wisdom of a people; language establishes identity and a sense of place and belonging.

British Columbia is unique in Canada for its multiplicity and diversity of Indigenous languages; there are eight (8) language families with 34 distinct languages, and 62 dialects. Each language has developed an orthography for its language. In addition, due to population dislocation, relocation, and dispersal, BC is also home to people from across Turtle Island representing their own languages Due to lack of access to resources each of the languages is in a different state of development; some have had many years of documentation, research, and protection while others have been limited in development due to limited human resources and inadequate financial support. Indigenous language experts have had to develop new strategies, knowledge and practices in recovering, revitalizing, and reclaiming their languages in a world dominated by the second language learning of colonial languages.

We are in a new colonial era, the new colonizers are those that are covering and infiltrating the world community with their vision of human connectivity. We the Indigenous peoples of the world can help give voice to understanding the experience of being colonized. Let me illustrate with some stories.

Our languages were mainly oral languages until the late 1960's. My story comes from that time, as my community wanted to teach our language in the school. At the time we used the typewriter, that technology drove how we designed our orthography. In our part of the country it is mainly the English language. With the introduction of the IBM Selectric typewriter our world changed when we found someone who could take the font ball and modify it – scrape off letters and symbols and add our letters, we could then type and produce our curriculum and story books in our language, we could record data and learn about our way of understanding the world. We could record and share with our community our knowledge of the plants, animals and the land, we could remember and share our ancestral stories. When computers came along we could no longer use that technology to do any of that work because our writing systems were not compatible. It took two 17 year olds, one in Ontario and another in Australia who figured out how our languages could access the new technology. It enabled First Voices to archive Indigenous languages, produce on line dictionaries, phrases, word games, and APPS. They created computer keyboards for every First Nations language in Canada, including syllabics, all the languages in Australia and the languages of the USA who chose to write their language, there are Indigenous peoples who refuse to shift their language from an oral language to literacy, as it changes the language and relationships. The downloadable keyboards enabled our people to communicate in our languages on line. Our languages nearly disappeared because it was promoted that in order to live in this world we had to only work in the colonizers languages. Technology today is still following that pattern.

One of the most powerful ways of eradicating the use of Indigenous languages, silencing the knowledge and wisdom of the people, breaking the intimate relationship with the earth, the lands and all that live there, was to remove people from their homelands and communities. It was accomplished by every one of the settler institutions – Education being the strongest force, removing children from their parents and grandparents, aunties and uncles their primary teachers to be educated away from their homes and communities, schooled in the settler languages; Social and family services, removing children from their families and communities to be raised outside of their communities away from their homelands and languages, note more Canadian Indigenous children are separated from their families than during the Residential school era; legal system, Indigenous men and women outnumber any other population in incarceration; family breakdown and violence results in women and children needing to leave their communities to find shelter, support, employment away from their home communities, As a result Indigenous people live away from their home communities often in large urban areas, some are not even in Canada, but in countries around the world. Technology would be the greatest tool to connect people to their languages but due to lack of technical expertise, connectivity, planned obsolescence and incompatibility of hardware and programming it is a challenge for Indigenous people in their language homelands to serve both those within their communities and those who live away.

Another example of the use of technology to reconcile Indigenous peoples knowledge system is in the area of mapping. When Indigenous peoples view maps either they don't exist or often they have been named by others – neighbouring Indigenous peoples, anthropologists, linguists, settler governments,

explorers and ethnographers. Names, boundaries and spelling are not the names, boundaries and spelling by the Indigenous people themselves. The tools people now have are enabling them to map their territories, include the songs and stories associated with each place, information about the histories of the location, what plants and animals thrived their and the unique words that were born in that place.

In a recent Environmental Scan survey of Indigenous language education in provincial schools showed that public school districts have the hardware and internet access even in remote areas but only one region of the province used technology for access to learning an Indigenous language. And in that one district they reported that students preferred face to face class learning of their Indigenous language and didn't find learning through technology helpful. In that region there are fewer Indigenous students enrolled and they speak many different Indigenous languages. More effort needs to be made to developing and implementing Indigenous language teaching and learning using distance technology.

It is only through working together in a respectful, collaborative way, in mutuality that we can help to keep Indigenous languages, thoughts, wisdom and nurturing relationships in this world and in our memory. It is our collective responsibility to keep our human communities healthy, to care for the earth, the water, the plants and animals. What is the role of technology in that care?

Bibliography

First Peoples Culture Council, (2018) (*Report on the Status of B.C. First Nations Languages 2018*

Galla, C.K. (2016) Indigenous language revitalization, promotion, and education: function of digital technology, Computer Assisted Language Learning, 29:7, 1137-1151, DOI: 10.1080/09588221.2016.1166137

Surma, A.; Truong, C.L. (2019) Digital Tools for Language Revitalization. https://abtec.org/iif.

Proceedings, pages 311–314
Paris, UNESCO Headquarters, 5-6 December, 2019.

21st Century Language Technology Tools – 21st Century Challenges vs. 21st Century Opportunities

Antti Arppe[1] & Jordan Lachler[2]
[1]Alberta Language Technology Lab, [2]Canadian Indigenous Languages and Literacy Institute, University of Alberta
arppe@ualberta.ca, lachler@ualberta.ca

Abstract

This paper presents a brief overview of the historical and current circumstances of Indigenous languages spoken in Canada, forming the basis for contemporary needs, challenges as well as opportunities presented in the development of modern language technological tools and applications for these languages in the 21st century.

Keywords: Canadian Indigenous languages, Algonquian languages, Dene languages, Cree, Intelligent on-line dictionaries, Corpora, Spell-checkers, Intelligent Computer-Aided Language Learning Applications

Tiivistelmä (in Finnish)

Tämä artikkeli esittää tiiviin yleiskatsauksen Kanadassa puhuttavien alkuperäiskielten olosuhteiden historiallisista kehityskuluista ja nykytilanteesta. Näiden perusteella artikkelissa kuvataan, mitä tarpeita, haasteita ja mahdollisuuksia on modernin kieliteknologian kehittämisessä näille alkuperäiskielille 21. vuosisadalla.

1. Introduction and Context

1.1 Indigineous Languages Spoken in Canada

Canada has much more linguistic diversity than e.g. Europe. In the 2016 Census of Population in Canada (Statistics Canada, 2017), participating people reported over 70 Indigenous languages, grouped into 8 distinct language families, namely the Algonquian, Inuit, Dene (Athabaskan), Siouan, Salish, Tsimshian, Wakashan and Iroquoian language families, as well as the isolates Kutenai and Haida, plus Michif which combines both Cree and French origins.

The aforementioned language families do not know national boundaries, spanning the Canadian-American and provincial/state borders**.** Dene languages are spoken from Alaska into Canada and hopping into South-Western United States; Algonquian languages are spoken on both sides of the border from the Rockies through the Great Lakes to the Atlantic, as are Iroqoian languages on both sides of Lake Erie and the St. Lawrence River, Siouan languages stretch north across the Western Plains, Salishan languages extend from British Columbia to the states of Washington, Idaho and Montana, and Inuit languages span the entire North-American Arctic from Alaska across Northern Canada to Greenland.

As many as 260,550 people in Canada, a number which has grown since 2006, reported being able to speak an Indigenous language well enough to carry out a conversation, out of 1,673,785 people (4.9% of the entire Canadian population) reporting Indigenous/Métis/Inuit identity/heritage**.** Nearly 213,225 people reported speaking an Indigenous language as a mother tongue, defined as the first language learned at home in childhood and still understood. Thus, the number of people able to speak an Indigenous language exceeded the number who reported an Indigenous mother tongue. According to Statistics Canada, this suggests that many people, especially young people, are learning Indigenous languages as second languages. (Statistics Canada 2017)

Algonquian languages were the family with the largest number of speakers, 175,825, in Canada, of which Cree with 96,575 speakers and Ojibwe with 28,130 speakers were the largest individual languages in this family. The next largest Indigenous language families and languages in terms of speakers were the Inuit languages (42,065 speakers, of which 39,770 spoke Inuktitut), the Dene (Athapaskan) languages (23,455 speakers, of which 13,005 spoke Dene [sųłiné]), followed by the rest, for which the numbers of speakers range from several thousand down to only a few tens. In terms of age groups, older Indigenous people were more likely to be able to speak an Indigenous language than younger generations, so that 35.6% out of seniors (65 years and older) could speak an Indigenous language, with the proportion decreasing with each subsequent age bracket. However, since there are four times more Indigenous children than seniors, in absolute terms there are more Indigenous children (45,135) than seniors (22,125) who could speak an Indigenous language. (Statistics Canada, 2017).

Furthermore, the results of this Census indicate that there were many times more Indigenous languages than Indigenous communities (over six hundred Sovereign First Nation communities) in Canada. Quite often and not surprisingly, most communities consider the language as it is spoken in that community as distinct and as a symbol of identity. In the Province of Alberta alone, there are nine Indigenous languages, which are (Plains or Woods) Cree, Blackfoot and Saulteux in the Algonquian family, Dene [sųłiné], Beaver, Slavey and Tsuut'ina in the Dene family, and Stoney/Nakoda in the Siouan family, plus Michif, spoken in as many as 46 First Nations Communities.

1.2 Historical Context in Canada

One could argue that language technology has hitherto generally proved to be of most use in supporting written language, though recently speech technology has finally also started to show real promise. In this respect, writing systems were developed and used for many Indigenous languages in Canada by the 1800s, for instance Aboriginal Syllabics for Cree, Blackfoot, and Ojibwe as well as Inuktitut (Rogers 2005). Furthermore, there has been a tradition of rapid uptake of Indigenous literacy. For instance, there is evidence of high literacy among Indigenous peoples in the 1800s, sometimes anecdotally claimed to have been higher than the proportion of literate people among the colonizing population (Rogers, 2005). What appears to have changed the general dynamics is Canadian Confederation in 1867 and the notorious Residential School system that it continued, which had started in the 1800s and lasted as long as until 1990s in some parts of Canada. This entailed the forced removal of many Indigenous children from their families to residential schools, often far away from their home communities, even in a different provinces, with children from multiple Indigenous language communities intermixed. In residential schools, the use of Indigenous languages by pupils was forbidden. Nevertheless, many children spent summers and other breaks with their families in their home communities, and returned there after finishing residential school, continuing to retain knowledge of their Indigenous mother tongues.

A possible turning point in the 1970s was the arrival of mass media, in particular TV, in many Indigenous communities. This resulted in significantly expanded exposure to majority languages (i.e. English and French in Canada) in previously isolated communities. Indeed, some members of Indigenous communities have personally noted the advent of mass media as the moment when their Indigenous language went into decline.

All in all, as a result of these aforementioned policies and technological changes, as well as other factors, the use of Indigenous languages was gradually squeezed into the oral realm, and more and more out of the public sphere.

1.3 Characteristics of Indigenous Languages in Canada and Their Resources

Many Indigenous languages spoken in Canada and the rest of North America are quite unlike majority languages such as English, in that they exhibit extensive morphological complexity, to the degree that there is no way to enumerate all possible word forms, or even all likely word forms, as one could do for English. In this respect they are similar to some other Indigenous languages, such as the Sámi languages spoken in Northern Scandinavia, for which extensive language technological tools have been created over the last decade. Table 2 illustrates with a Plains Cree text passage the extent of morphological complexity in practice, in a language with extensive morphological complexity, where only 10 of the 21 Cree words are easily listable non-inflecting forms, and 4 rarer inflected forms that would not normally be included in any core or even extended inflectional paradigm. Thus, without computational modeling of inflection and word structure, less than half of the words in running text could be matched *as is* with a dictionary entry, meaning that over half of the words would remain inaccessible using a software application that would not make use of a computational morphological model (cf. Klavans 2018).

ᐃᖏᐦᐋˣ ᑲ ᑭ ᐊᐋᓯᓱᐃᓵᐣ ᓂᑭ ᓂᑕᐃ ᑭᖏᑭᓇᐦᐊᒪᓱᐣ ᐅᖏᐱᒪ ᓂᑎᐦᑯᓂᑲᓂᓂˣ ᐅᖏᐸx ᑭᖏᑕᐱᓀᓂˣ ᒪ ᓀ ᑲ ᑭ ᓂᑕᐃ ᑭᖏᑭᓇᐦᐊᒪᓱᐣ, ᐃᑯᑕ Residential School ᐁ ᑭ ᐃᑐᐢᑕᐢᐃᑲᐃᓵᐣx ᒥᖏᑕᐢᐃ ᐊᓀ ᓀᒧᓵ ᐁᐊᐧ ᑭᑎᒪᑭᐢᐃᑲᐃᓀᓵᐣ, ᓀᑎ ᐊᑎᒪᐊᐧ, ᓀᑎ ᐅᑭᖏᓇᐦᐊᒪᓀᐠx

Table 1. The 1st paragraph of a Plains Cree story, *Dog Biscuits* by Solomon Ratt, in (Western) Canadian Syllabics, courtesy of Cree Literacy Network (URL: https://creeliteracy.org/2014/01/20/dog-biscuits-y-dialect-with-audio/). Words in **red** are non-inflecting words; words in **blue** are complex inflected forms outside any core paradigm, and words in **black** inflected forms in the core paradigm.

Moreover, written corpora, if any exist, are relatively small, with at most several hundred thousand word form tokens representing 10-20 thousand word form types (e.g. Arppe et al., forthcoming, in the case of Plains Cree). Thus, these corpora, in terms of their representativeness of word form types, are not conventionally sufficient for applying machine learning to learn the languages' morphological system. On the other hand, there often exist relatively comprehensive lexical databases as well as complete inflectional paradigms for many of these Indigenous languages, as a result of extensive linguistic documentation work in the 19th and 20th centuries. Importantly, these are resources that are highly amenable for 20th century rule-based computational modeling techniques of morphology using finite-state transducers (cf. e.g. Beesley & Karttunen 2003), see e.g. Harrigan, Schmirler, Arppe et al. (2017) in the case of Plains Cree.

2. Challenges

As a result of the historical context and developments or accidents, there exist few if any established, obvious institutions for collaboration concerning Indigenous languages in Canada, which typically would concern multiple communities speaking the same language or its dialects across several provinces (e.g. Cree or Ojibwe). This has been facilitated by the splintering of the language communities through the reserve system. Furthermore, one can observe a predominance of oral tradition, though we should recall the history of Indigenous literacy in the 1800s. This has been bolstered by the fact that there are few large collections of written texts or literature, as noted above, which can generally be associated with people having less exposure to written materials in their heritage language.

Moreover, orthographical standard forms are sometimes seen as representing a continuation of colonial thinking. Alongside this view, one can also see the success of the principle "The language is written as it is spoken", resulting in substantial variation in how the languages are written, even from one individual to another. On the other hand, this lack of standard also allows for representing the rich variation from individual/community to another. Nevertheless, common orthographical conventions would allow others to more easily both find and understand what has been written by someone else.

Another set of challenges arises from representing appropriately sounds not apparent in the majority languages of Canada, English and French, in a native way. Historical, pre-computer era solutions to this have included using underlining as a diacritic (e.g. Haida), the adoption of apostrophes for glottal stops (e.g. Haida) and other punctuation marks or numerals for non-Western phonemes (in particular for Indigenous languages in British Columbia), and marking e.g. long vowels or geminate consonants as in IPA using a colon (e.g. Mohawk). Furthermore, one can sometimes observe the explicit marking of allophonic, non-phonemic differences in pronunciation, which can result in unnecessarily detailed, complex writing systems. All this poses in the contemporary world the practical challenge of how to to deal with and input some of such non-English characters for non-English sounds on computers and mobile devices. Nevertheless, to some extent, this can be addressed with the flexible expandability of character sets and scripts within the Unicode standard, fuzzy matching algorithms linking what a user types with the properly spelled word form using the standard characters according to a language's writing system, and clever keyboard design.

Moreover, one can observe not only multiple scripts, but also multiple orthographical standards/conventions per these scripts, for writing the very same utterance, as a result of the historical context and developments and accidents. Only for (Plains) Cree (Table 2), one can, or has to, choose between Canadian Aboriginal Syllabics vs. Standard Roman orthography (SRO). Furthermore, there are varying conventions as to whether and how one marks, or does not mark, long vowels (with circumflexes or macrons or not at all) and syllable final aspirations (*-h*) with diacritics. In addition, there is variation, in both Standard Roman Orthography (SRO) and Canadian Syllabics, concerning whether one writes prefixes or suffixes joined with the stems (with or without a hyphen), or not, separating the affixes with spaces.

kâ-kî-awâsisîwiyân
kā-kī-awāsisīwiyān
kâ kî awâsisîwiyân
kakihawasisiwiyan
ᑳ ᑭ ᐊᐋ·ᓯᓯᐃᐏᔭᐣ

Table 2. Varying ways of writing the (Plains) Cree word *kâ-kî-awâsisîwiyân* 'when I was a child'

This poses a general challenge for language technological tools and applications of recognizing as well as generating these multiple standards and conventions in writing. Nevertheless, this is in principle quite straight-forward to address with the computational modeling of both word structure and of orthographical variation, which allows the easy creation of transcriptors converting between all/known, standardized variants.

3. Expectations of Indigenous Communities and Individual Speakers and Learners

Based on our experiences, Indigenous communities in Canada we have worked with have appeared to value recording the richness of their vocabulary, as known by Elders and other fluent native speakers, and also recording how their language is properly spoken by Elders; in contrast, the creation of proofing tools to support the proper writing of texts, while potentially valued, has not always been considered as high in urgency. Indigenous learners/speakers have appeared to value the ability to look up words in an Indigenous language for a concept (i.e. an English word), the ability to find out what a particular word in an Indigenous language means (i.e. a translation into English), and the ability to write, or often importantly say, a particular word or phrase in their Indigenous language. Nevertheless, one needs to recognize here that there are in fact multiple Indigenous speaker/learner subgroups within even a single First Nations community, e.g. Elders, other fluent speakers, language instructors/teachers, school pupils and other language learners, as well as other community members. These different subgroups have different levels of language proficiency, different linguistic community contexts, and consequently different needs and priorities; thus, they also need, expect and would benefit from different language technological solutions. Furthermore, these subgroups within individual First Nations communities may have more in common with similar subgroups in other communities speaking the same Indigenous language, perhaps even more so than with other subgroups in their own community.

As for language technological applications and Indigenous communities, electronic/on-line dictionaries are often considered very desirable, and having a spoken component included in such a resource is considered very important. Morphological intelligence, that is the recognition and generation of complex word-forms is certainly seen as a plus, but this feature is not among the first expectations. Moreover, there is a concern that word-form/paradigm generation produces non-validated unnatural or weird forms. Furthermore, proofing tools such as spell-checking or predictive text do not appear to be the very first expectation, though views of community members seem to change when shown demo mock-ups illustrating such a functionality for their own Indigenous language.

As a language-independent but relevant recent development, the last decade has witnessed a diffusion of mobile/digital devices; Indigenous youth have become digital natives (and increasingly older folks, too). The youth are often considered the "missing generation" in language revitalization, since they did not learn the language at home, but are now too old to benefit from language nests and immersion schools targeting younger

children. Indigenous teens and young adults are indeed a vital piece of the puzzle in language revitalization. We are seeing an increasing amount of mobile media use which is based on written communication, and with this the emergence of need of language technology supporting written language, where spell-checking, predictive text, and word-form generation based on computational modeling has a role. Moreover, Indigenous individuals are more and more moving and living outside the reserve in urban centers, without direct access to Elders and other fluent speakers (most of whom are still remaining on the original reserves); thus, they would greatly benefit from written and spoken Indigenous language resources available on-line.

4. Opportunities and Solutions

The development and diffusion of digital devices makes literacy-based language tools useful for a critical generation of Indigenous communities in Canada, as well as others. In this, we in the Alberta Language Technology Lab (ALTLab: altlab.artsrn.ualberta.ca) have been inspired by what the Giellatekno and Divvun research and development teams at UiT – Arctic University of Norway have been able to create for the Indigenous Sámi languages, and have started adapting their work to the Canadian context, encountering both similarities and differences in the circumstances of Indigenous languages in Canada. Following their example, we have aimed at the "low-hanging fruit" that can be created with the existing scarce but rich documentation resources which are amenable to rule-based computational models, and language technological applications and resources that can be created with such models (Arppe et al., 2016). These include (1) **web-based intelligent dictionaries** (I-DICT) presenting both the written and spoken form of words, and that allow for searching with inflected forms and the generation of inflectional paradigms (altlab.ualberta.ca/itwewina); (2) **searchable databases** of both written and spoken usage examples (Arppe et al., forthcoming: altlab.ualberta.ca/korp); (3) **spell-checkers** to support the creation of high-quality texts by speakers and learners; and (4) **intelligent computer-aided language learning** (I-CALL) applications which include training in both the spoken and written forms of the language (Bontogon et al, 2018: oahpa.no/nehiyawetan). To date, we have created our first full demonstration versions of these tools only for Plains Cree, but have started work on similar tools and applications for several other Indigenous languages spoken as well in Canada. Importantly, though such tools are oriented firstly towards supporting literacy, it is worth noting that they can also provide substantial support in creating spoken resources and tools, thus presenting a further significant benefit as oralcy is valued highly by many Indigenous communities.

Alongside our work, we need to note a substantial number of parallel on-going projects on developing language technology for Indigenous languages in Canada, a comprehensive overview of which is presented in Littell et al. (2018).

5. Conclusion

The recent development and diffusion of digital devices makes literacy-based language tools useful for a critical generation of Indigenous language learners and speakers, And indeed, supporting literacy is historically not entirely foreign to Indigenous languages. In all this, we consider serving the needs and expectations of Indigenous communities as paramount, since we want our tools to be of genuine use to these communities, but we also recognize that there are in fact multiple user subgroups whose needs may diverge.

9. Acknowledgements

We acknowledge funding by the Social Sciences and Humanities Council of Canada (awards #890-2013-0047, #611-2016-0207, and #895-2019-1012), and the University of Alberta, Kule Institute of Advanced Study (KIAS Research Cluster grant 2015-2018).

6. Bibliographical References

Arppe, A. Lachler, J., Trosterud, T., Antonsen, L. and Moshagen, S. N. (2016). Basic Language Resource Kits for Endangered Languages: A Case Study of Plains Cree. *CCURL 2016 - Collaboration and Computing for Under-Resourced Languages – Towards an Alliance for Digital Language Diversity*, 1-8, Portorož, Slovenia, 23 May 2016. URL: http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-CCURL2016_Proceedings.pdf

Arppe, A., Schmirler, K., Harrigan, A. G. and Wolvengrey, A. (forthc.). A Morphosyntactically Tagged Corpus for Plains Cree. *Papers of the 49th Algonquian Conference*, Oct. 2017, Montréal, Quebec.

Beesley, K. R. and Karttunen, L. (2003). *Finite-State Morphology*. California: CSLI.

Bontogon, M., Arppe, A., Antonsen, L., Thunder, D. and Lachler, J. (2018). Intelligent Computer Assisted Language Learning (ICALL) for *nêhiyawêwin*: An In-Depth User Experience Evaluation. *Canadian Modern Language Review*, 74(3), 337–362

Harrigan, A. G., Schmirler, K., Arppe, A., Antonsen, L., Trosterud, T. and Wolvengrey, A. (2017). Learning from the computational modelling of Plains Cree verbs. *Morphology*, 27(4), 565–598.

Klavans, J. L. (2018). Computational Modeling of Polysynthetic Languages. Proceedings of Workshop on Polysynthetic Languages, pages 1–11 Santa Fe, New Mexico, USA, August 20-26, 2018.

Littell, P., Kazantseva, A., Kuhn, R., Pine, A., Arppe, A., Cox, C. and Junker, M-O. (2018). Indigenous language technologies in Canada: Assessment, challenges, and successes. In Bender, E., Derczynski, L. and Isabelle P. (eds.), *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, 2620–2632. Santa Fe, New Mexico: ACL.

Rogers, H. (2005). *Writing systems: a linguistic approach*. Blackwell publishing.

Statistics Canada (2017). Census of Population, 2016. URL: https://www12.statcan.gc.ca/census-recensement/2016/as-sa/98-200-x/2016022/98-200-x2016022-eng.cfm

Proceedings, pages 315–319
Paris, UNESCO Headquarters, 5-6 December, 2019.

Indicators of Languages in the Internet

Daniel Pimienta
Observatory of Languages and Culture in the Internet
http://funredes.org/lc
World Network for Linguistic Diversity
http://maaya.org

Abstract

The availability of indicators of the space of languages on the Internet is required to support appropriate public policies. Current sources are scarce and strongly biased. This approach computes indicators for the 140 languages with more than 5 million L1 speakers. It relies on the collection of a large set of micro-indicators measuring languages or countries in various Internet spaces or applications. Statistical methods are applied to produce 6 indicators: Internet users, traffic, use, contents, societal indexes and interfaces, from which 4 macro-indicators are deduced: power, capacity, gradient and content productivity. Some results are presented and the biases of existing methods are analyzed.

Keywords: Languages, Internet, Indicators, Biases

Résumé

Il est nécessaire de disposer d'indicateurs de la place des langues dans l'Internet pour pouvoir conduire des politiques publiques. Les sources disponibles son rares et fortement biaisées. Cette approche calcule des indicateurs pour les 140 langues de plus de 5 millions de locuteurs L1. Elle s'appuie sur la collecte d'une large série de micro-indicateurs mesurant les langues ou les pays d'une variété d'espace ou d'applications de l'Internet. Des méthodes statistiques sont appliquées pour produire 6 indicateurs : utilisateurs de l'Internet, trafic, usages, contenus, index sociétaux et interfaces, à partir desquels 4 macro-indicateurs sont déduits: puissance, capacité, gradient et productivité de contenus. Quelques résultats sont présentés et les biais des méthodes existantes sont analysés.

1. Introduction

During the period 1998-2007, the Observatory of Languages and Cultures in the Internet[1] has been a project of Networks & Development Foundation (FUNREDES[2]) and has collaborated with Union Latine[3] for the design of methods for measurement of language's in the Internet which could provide reproducible and reliable indicators; at the same time other initiatives[4] existed with the same objectives. (Pimienta, 2009). From 2007, changes in the size of the Web and search engines behaviors has rendered obsolete the methods and created a vacuum in the production of indicators of languages in the Internet. A new *artisanal* method, based on the observation of language's behavior in a wide variety of spaces and applications of the Internet was proposed in 2012 and opened new studies of the Observatory, under the World Network for Linguistic Diversity[5] institutional hat and with the support of OIF[6]. Two early studies provide results in terms of rankings for French in the Internet. The second, conducted in 2013, fed the Internet chapter of the 2014 report "Le français dans le monde" (OIF, 2014) and was followed by a similar study of Spanish in the Internet (Pimienta D., Prado D., 2016). The latest OIF funded study, more ambitious, which inspires this article, managed, by the application of a statistical approach, authorized by the increased number of sources, to achieve results in terms of language indicators in the Internet for a wide range of languages.

The method is based on collecting quantitative information about language use in as many as possible applications and Internet spaces. The statistical process of sources enable the measurement of the presence of languages in the Internet and put the results into perspective by building a series of indicators of the share of languages in the Internet. A synthesis is extracted in the form of a series of macro-indicators which combine all indicators. The methodological framework is to use sources either directly when figures concerning languages are available, which is unfortunately rare, or indirectly, using figures per country and transforming them into figures per language. This transformation makes this method an unprecedented approach with the ability to handle the language data quest, in a context where language indicators have become, at best, highly unreliable, but mostly and usually nonexistent.

[1] http://funredes.org/lc
[2] http://funredes.org
[3] http://unilat.org
[4] In particular the ambitious Language Observatory Project (Mikami et al. 2006)
[5] http://maaya.org
[6] http://francophonie.org

This approach is supported by implicit assumptions that need to be made explicit and evaluated to ensure consistency, reliability and expose the corresponding biases. The results are compared with the 2 existing sources: (W3Techs, 2019) and (InternetWorldStats, 2019) and the notable differences are analyzed under the focus of the respective biases.

The details of the methodology, a compilation of the results and the complete biases analysis of the 3 existing methods can be consulted in (Pimienta, 2017) and in http://funredes.org/lc2017.

2. Indicators

The following diagram shows all the indicators which are processed for each language and the corresponding quantity of sources..

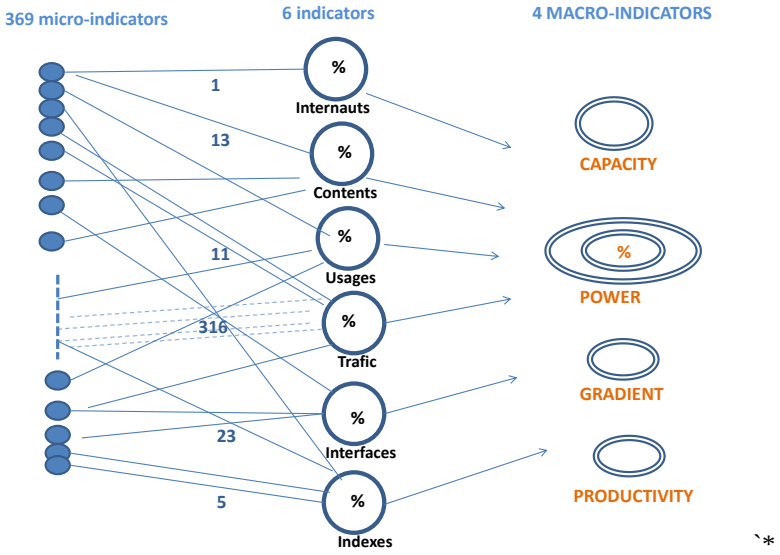


Figure 1 : Indicators diagram

The following table shows for each indicator its sources and how it is computed. All the indicators are expressed in terms of world share, based on the total population of L1+L2 speakers.

INDICATOR	DEFINITION	PROCESS	RELIABILITY
A: INTERNAUTS	Single indicator from ITU: world % of people connected per country.	Weighting country -> language	Very strong Only marginal bias
B: USES	Includes 11 micro indicators: Telephone lines; e.commerce market; OpenOffice download; Social networks users+2021 projection; various social networks subscribers and projections	Weighting C-> L Extrapolation by proportion Truncated mean at 20%	Strong reliability. Low bias. But the number of micro-indicators would need to be extended to give more sense to the mean.
C: TRAFFIC	Alexa.com measured traffic from a selection of 316 websites.	Weighting C-> L Extrapolation proportion. Truncated mean to 20%	Relatively good But huge occidental bias of Alexa
D: INFORMATION SOCIETY	Includes 5 indexes from WebIndex for the following criteria: E.gov, Universal Access, E.participation...	Weighting C -> L, mean Transform to world % by weighting with ITU data	Good (subjective data by competent body). Should be extended.
E: CONTENT (Wikipedia and books)	Includes 13 micro indicators Number of books at Amazon; W3Techs; 11 language indicators from Wikimedia	Direct use of figures per language. Truncated mean to 20%	Very strong. But strong negative bias for Asian languages. Need to be extended
F: **INTERFACE** (and translation languages)	23 binary micro indicators 12 interfaces, 1 content language, 10 translation applications	Presence % transformed in word % by weighting with ITU figures.	Perfect.

Table 1 : Indicators description

The following table shows the defined macro-indicators.

POWER	Measures the global share of the language in the Internet	Mean of the 6 indicators (world L1+L2 %).
CAPACITY	Measures the strength of the language in the Internet regardless of its number of speakers.	Ratio of power vs. world % of speakers. No dimension, normalized to 1.
GRADIENT	Measures the strength of the connected speakers regardless of their number.	Ratio of power vs. world % of connected speakers. No dimension, normalized to 1.
PRODUCTIVITY	Measures the propensity of the connected speakers to produce content in their language.	Ratio of % of contents vs. % of connected speakers. No dimension, normalized to 1.

Table 2 : Macro-indicators description

3. Computations

The model stands on a 3 categories of data : 1) the large list of Internet related sources for language or country 2) Demo-linguistic data 3) ITU data the percentage of people connected to the Internet per country (ITU, 2009).

1) All micro-indicators are expressed into world percentage. The transformation from country to language is realized by weighting with the number of speakers of each language in each country. The sources rarely cover all countries in the world then some extrapolation techniques are used, either in proportion of the percentage of people connected by country or using the method of quartiles. Whenever the extrapolation lacks sense the micro-indicator is rejected.
2) Two sources exists as of today providing the matrix of quantity of L1 speakers of each language for each country: the Joshua project (free of charge) and Ethnologue (fee required). The first edition of the model used Joshua. As for L2 speakers Ethnologue data is used and future measurements will try to use Ethnologue for L1 as well.
3) ITU data, regarded as both reliable and essential to the method, is updated free of charge each year.

The LOC_1 matrix meets the followings definition, for all selected languages and all selected countries: $LOC_1(i,j)$ = Number of L1 speakers for the language i in country j.

The source provides figures for 7500 languages but only a subset will be processed. The estimated number of languages present in the Internet is around 500. One possibility is to target them. Another possibility is to select the languages for which Wikipedia offers statistics (close to 300). After several tests the choice finally settled on the list of the 140 languages with more than 5 million speakers The decision was made in order to reduce the biases resulting from the implicit assumptions.

As for L2, the first priority is that of taking coherent account of multilingualism. The persons computed in L2 obviously have also a first language and therefore the set of L1+L2 speakers includes the same persons more than once. The evidence says that figures must necessarily be based on the total language speakers in the world and not in relation to the world population. This evidence is unfortunately ignored by many sources and provokes errors. In the scenario that is adopted the world share will be calculated on the basis of 125% of the world population (figure computed from the demo-linguistic inputs). This notion is equally applicable to all concepts: users, traffic, usage, content, interfaces and indexes (for instance websites can be made in several languages, the same for the flow of emails). The ideal method to treat the case of L2 would obviously be to produce, as for L1, a matrix $LOC_2(i, j)$ = number of L1+L2 speakers of language i in country j. Unfortunately, this data is unavailable. It is then proposed another approach which simple principle consists, for each language, to get a number that represents the increase to be applied to L1 quantities to get the value L1+L2 and use a linear approach for the results. The global rate of increase (1.25) is the result of the following weighting operation:

$$Rg = \sum_{j=1}^{j=L} L_1(j) \times R_{12}(j)$$

where L is the total number of languages, $L_1(j)$ the number of L1 speakers for language j and $R_{12}(j)$ the rate of increase from L1 to L1+L2 for the language j. The value of micro indicators for L1+ L2 are calculated this way from the value for L1:

$$M_{L1+L2}(i) = Rg \times M_{L1}(i) / \sum_{j=1}^{j=L} M_{L1}(j) \times R_{12}(j)$$

The L1+L2 method is applied to all indicators except Index, Content and Interface which are by nature meant to apply directly to L1+L2. This

method is less accurate than a solution that could work at the country level and generate some biases.

As for computing the indicators, only those expressed by countries require computation. The principle to converts figures expressed in percentages per country into percentages per language is the matrix product between the LOC matrix and, the vector MCn containing the source figures per country for micro-indicator n. The micro-indicator expressed in percentage per language (MLn) is then:

$$MLn(i) = \sum_{j=1}^{j=P} LOC(i, j) \times MCn(j)$$

where P is the total number of countries, LOC(i, j) is the number of speakers of language i in country j and MCn(j) is the measured value for the micro-indicator n in country j.

The matrix product ML = LOC +. x MC in APL[7] notation or ML=SumProduct (LOC;MC) in Excel notation, is a weighting operation of the values of the micro-indicator in each country with the presence of each language in each country. The MLn's totals are the same as those of MPn but this time the distribution is made per language instead of per country. As most of the computations are based on weighting it is useful to identify the different types used in the process and make explicit the simplifying assumptions underlying the validity of the results obtained by these weightings, assumptions which will guide the understanding of biases.

	Demo-linguistic	L2	Users
TYPE	C ---> L	L1 ---> L1+L2	Criterion % -> world %
APPLICATION	Data by C	L1 Results	% by criteria
RESULT	Data by L	L1+L2 Results	% worldwide
WEIGHTING	LOC matrix	L1+L2/L1 per L	IUT data
SCOPE	All sources by country	Users, traffic and usage	Index and interfaces
IMPLICIT ASSUMPTION	Identical connection rate for all L1 in the same C	Identical connection rate for all L2 as for L1	Modulation according to Internet connection rate

Table 3: Different weightings applied

[7] APL, "A Programming Language", a mathematical formalism and programming language.

4. Results

The following table shows the bias corrected results for the first 10 languages for contents and allows comparison with the two other existing sources, showing strong discrepancies which are well understood when the biases are analyzed carefully (Pimienta, 2017).

	CONTENTS	W3TECH	INTERNAUTS	IWS
English	**32,0%**	51,9%	20,4%	26,3%
Chinese	**18,0%**	2,0%	20,0%	20,8%
Spanish	**8,0%**	5,1%	9,1%	7,7%
French	**6,5%**	4,1%	4,9%	2,8%
German	**3,8%**	5,5%	2,7%	2,3%
Portuguese	**3,5%**	2,6%	4,1%	4,3%
Japanese	**3,5%**	5,6%	4,5%	3,2%
Russian	**3,5%**	6,5%	4,9%	2,9%
Hindi	**3,0%**	< 0,1%	4,6%	n.a.
Arabic	**3,0%**	0,7%	3,0%	4,7%
Remaining	***40,2%***	*15,9%*	*46,6%*	*25,0%*
TOTAL	125,0%	100,0%	125,0%	100,0%

Table 4: Ten top languages content & comparisons

The following table shows the ranking for capacity and it is not surprising to see the languages of countries with strong policies for Information Society.

	Capacity	Ranking Power	% connected
Hebrew	5.40	35	76.05
Finnish	5.40	38	92.30
Dutch	4.81	19	92.27
Swedish	4.46	28	90.54
English	3.72	1	78.05
German	3.40	6	86.43
Danish	3.30	49	95.67
Italian	3.16	12	64.20
Czech	3.13	27	81.17
French	2.96	4	81.09

Table 5 : Ten top languages for capacity

And finally the table, sorted by gradient highlights the dynamism of people connected.

	Gradient	Ranking Power
Hebrew	2.62	35
Finnish	2.16	38
Dutch	1.93	19
Swedish	1.81	28
English	1.76	1
Czech	1.42	27
English	1.76	1
Italian	1.73	12
Serbo-Croatian	1.54	22

Table 6: Nine top languages for gradient

5 Bibliographical References

Ethnologue, (2019). Languages of the World. https://www.ethnologue.com

Internet World Stats, (2019), Internet world users per language, top 10 languages. https://www.internetworldstats.com/stats7.htm

ITU, (2019), Percentage of individuals using the Internet per country. https://www.itu.int/en/ITU-D/Statistics/Documents/statistics/2019/Individuals_Internet_2000-2018_Jun2019.xls

Mikami Y., et al. (2005), The Language Observatory Project (LOP), In Poster Proceedings of the Fourteenth International World Wide Web Conference, 2005, pp. 990-991, May 2005, Japan

OIF, (2014), Le français dans l'Internet, *Rapport 2014 "La langue française dans le monde"*, pp. 501-541, Nathan. http://francophonie.org/Rapports-Publications.html

Pimienta D., (2017) An alternative approach to produce indicators of languages in the Internet in *Proc. of Global Expert Meeting Multilingualism in Cyberspace for Inclusive Sustainable Development*, Khanty-Mansiysk, Russian Federation, June, 2017 http://funredes.org/lc2017/Alernative%20Languages%20Internet.docx

Pimienta D., Prado D., (2016) Medición de la presencia de la lengua española en la Internet: métodos y resultados, en *Revista Española de Documentación Científica* 39(3), julio-septiembre 2016, e141- ISSN-L:0210-0614. doi:http://dx.doi.org/10.3989/redc.2016.3.1328

Pimienta, D., Prado D. et al, (2009), Twelve years of measuring linguistic diversity in the Internet: balance and perspectives, in *UNESCO Publications for the World Summit on the Information Society*, CI.2009/WS/1 http://unesdoc.unesco.org/images/0018/001870/187016e.pdf

W3Techs, (2019), Usage of content languages for websites. https://w3techs.com/technologies/overview/content_language/all

6. Acknowledgements

The idea to use various country sources and transform them into language data was first conceived by Daniel Prado in 2012.

The study was funded by OIF.

Proceedings, pages 320–323
Paris, UNESCO Headquarters, 5-6 December, 2019.

Challenges for Language Technologies in Critically Endangered Languages

Jhonnatan Rangel
INALCO, SeDyL
7 Rue Guy Moquet, 94800 Villejuif
jhonnatan.rangelmurueta@inalco.fr

Abstract

There are currently 577 critically endangered languages in the world, making up almost 10% of all languages. These languages are also technologically low-resourced and are only spoken by a few elder speakers. As such, critically endangered languages pose various fundamental challenges, such as the annotation bottleneck, that seriously hinder future perspectives of language documentation, preservation, reclamation, revitalization and utilization in language technologies. This paper addresses the challenges critically endangered languages face in implementing language technologies.

Keywords: low-resourced, critically endangered, language technology

Resumen

En el mundo hay 577 lenguas en muy alto riesgo de desaparición que representan casi el 10% de todas las lenguas. Estas lenguas, que además cuentan con pocos recursos tecnológicos, las hablan únicamente unos cuantos adultos mayores. Las lenguas en muy alto riesgo de desaparición plantean retos fundamentales, como el cuello de botella de anotación, que limitan enormemente las perspectivas de documentación, mantenimiento, recuperación, revitalización y uso de tecnologías del lenguaje. Este artículo aborda los retos que enfrentan las lenguas en muy alto riesgo de desaparición en cuanto a la implementación de tecnologías del lenguaje.

1. Critically endangered languages

There are currently 577 critically endangered languages in the world, making up almost 10% of the world's 6,000+ languages (Moseley, 2010).

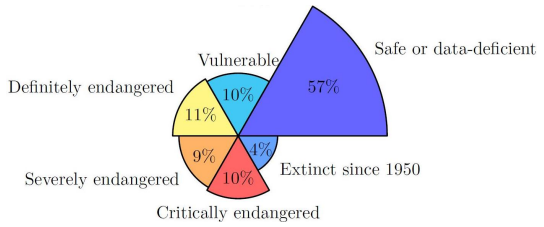


Figure 1: Vitality of the world's languages

These languages present the following characteristics that directly impact their vitality in the short term (Rangel, 2019):

- Youngest speakers are grandparents and older
- Small proportion of speakers in relation to larger community
- Inter-generational (or traditional) transmission of the language interrupted for decades
- Infrequent use of the language in language practices of remaining speakers
- Lack of written tradition and literacy in the language

Critically endangered languages can intersect with limited documentation and/or limited studies, and in most cases they are also indigenous or autochthonous languages. All of these characteristics can coincide such as in the case of Ayapa Zoque, Ayapaneco or *numde 'oode* (autonym), an indigenous, under-documented, under-studied and critically endangered language spoken in southern Mexico (Rangel, 2017; Rangel, 2019).

Because language endangerment is a global phenomenon, there are critically endangered languages on every continent of the world (Moseley, 2010). As these languages are at the highest level of endangerment, their disappearance could occur at any time in the next decade. Consequently, concrete and multifaceted measures must be taken immediately to reverse or at least slow down language endangerment in critically endangered languages such as Ayapaneco.

2. Low-resourced languages

Low-resourced languages (LRL) do not have the extensive resources required (annotated and parallel corpora) for the implementation of Language Technologies and techniques such as Machine Translation or Machine Learning. It is estimated that out of the world's 6,000+ languages, only about 20 of them have the resources to be considered high-resourced languages (HRL) while an additional 60 have some sort of resources available to be considered medium-resourced languages (MRL) (Duong, 2017).

HRL	MRL	LRL
0.4%	1%	98.6%

Table 1: Resource distribution of world languages

This means that Language Technologies are only applied to about 1.4% of the world's languages, leaving the vast majority of them unattended. Examples of HRL include English, Spanish or French while MRL include Hebrew,

Indi or Czech. These languages are spoken by millions of people of multiple generations in the world and therefore are not necessarily at immediate risk of disappearing.

On the other hand, some LRL are spoken by millions of people and are not at risk of immediate disappearance, such as Swahili. However, LRL can be critically endangered languages such as Ayapaneco. While there is a tendency for critically endangered languages to also be LRL, the opposite correlation does not always hold (not all LRL tend to be critically endangered languages). Indeed, a language's number of speakers does not necessarily determine the resources available. In fact, the development of language resources is strongly influenced by social, political, and financial factors. For instance, languages that are considered international and spoken predominantly in Western, Educated, Industrialized, Rich and Democratic (WEIRD) societies (Henrich et al., 2010) have an abundance of resources while minority languages spoken in non-WEIRD societies significantly lack resources (King, 2015).

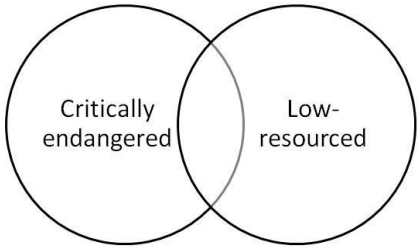


Figure 2: Linguistic and technological characteristics

I will address the challenges faced by languages situated in the overlapping zone between low-resourced and critically endangered that accounts for almost 10% of the world's linguistic diversity.

3. Challenges

Low-resourced critically endangered languages face multiple and multifaceted challenges for implementing language technologies. I will address 6 major challenges.

3.1. Annotation bottleneck

The annotation bottleneck corresponds to the gap between the large amount of data that we are capable of gathering with current technology and the limited amount of data we are capable of annotating (transcribing, translating and glossing). Estimates suggest that for every hour of recording, between 40 and 100 hours are required to annotate it (Seifart et al., 2018).

Although the annotation bottleneck is a challenge that can impact every language in the world, it is amplified in the context of critically endangered languages. The reason for this, as mentioned above, is that critically endangered languages are also LRL. Consequently, applying Natural Language Processing (NLP), Machine Learning or Artificial Intelligence tools and techniques that could help us to open up this annotation bottleneck is convoluted. By contrast, these tools and techniques can be implemented more easily in HRL to reduce the gap between data available and annotated data.

Although promising advances have recently been made in implementing Language Technologies for LRL (Ćavar et al., 2016), the annotation of critically endangered languages continues to be a primarily manual task carried out by a few researchers and community members, contributing in turn to the existing annotation bottleneck. This brings us to the second group of challenges, human resources.

3.2. Human resources

Human resources tend to be very scarce in the context of critically endangered languages. From the amount of existing speakers to the potential manual annotators, human resources are limited.

The potential universe of annotators of these languages is reduced to a language expert (which could be a linguist, an anthropologist or a teacher) and in best case scenarios, a few speakers. As mentioned, in the context of critically endangered languages, the youngest speakers are grandparents and older. Consequently, the remaining speakers of these languages cannot always contribute to the annotation process because as they get older, they suffer from physical conditions such as vision and hearing problems that prevent them from participating in the annotation of their language, not to mention that they also have very limited digital literacy.

A possible solution to this challenge is recruiting young community members who could help annotate data. The caveat is that in many critically endangered languages, the younger community members do not know the language well enough to perform this task by themselves and they need the help of elder speakers. As the number of remaining speakers is limited and their physical conditions are not optimal, this task becomes very slow and cumbersome.

I usually spend 60-120 hours annotating for every hour of Ayapaneco recorded using assisted annotation tools such as ELAN (Wittenburg et al., 2006). One younger community member with limited digital literacy assists with the annotation. However, neither of us know the language well enough to perform this task by ourselves and still require the speakers' input, resulting in very slow progress.

Recently, some tools based exclusively on oral annotations such as SayMore (Moeller, 2014) have emerged as a solution to opening up the annotation bottleneck. Although this can be a promising option for some world languages, it can be complicated to implement it in critically endangered languages because speakers are still required to contribute to the oral annotations, and speakers are scarce in these languages.

3.3. Capacity

As mentioned above, critically endangered languages are in most cases indigenous or autochthonous languages and

are also minority languages spoken in societies with poor economics. This directly impacts the capacity and the infrastructure available for these communities.

A very common capacity challenge among communities in which critically endangered languages are spoken is the scarce access to computers. Contrary to WEIRD societies, computer access can be very limited in the context of critically endangered languages as these communities most likely face poverty and marginalization. Without computers, it is difficult to introduce Language Technologies in these communities.

The recent global democratization of cellphones could facilitate the introduction of Language Technologies in these communities provided that access to internet is guaranteed. Unfortunately, this is not always the case. In Ayapa, the village where Ayapaneco is spoken, very few people own a computer but cellphone availability has dramatically increased in recent years with many community members owning one.

A second capacity challenge is related to literacy. In order for critically endangered languages to have some sort of online presence (ex: social media), they need to be written. However, as the case of Ayapaneco illustrates, not all world languages have a writing system, and when it comes to critically endangered languages, this seems to be the norm rather than the exception. Indeed, orthography development in Ayapaneco is a recent endeavour, and the writing system is not yet functional. Currently only two people are familiar with the orthography and therefore Ayapaneco is not yet used online.

Most speakers of critically endangered languages tend to be bilingual in their minority language and the majority language of the wider society. Nevertheless, when speakers of critically endangered languages are literate, they are only so in the majority language. Consequently, there is a lack of written tradition and literacy in these languages, complicating the task of applying Language Technologies.

3.4. Infrastructure

Closely related to the previous point, internet access tends to be a common challenge among endangered language communities. When internet access is available, it can be expensive to access given the poverty and marginalization discussed above. Critically endangered languages are commonly spoken in rural areas with limited or unreliable internet access. In Ayapa, community members access the internet mainly via mobile internet with a Smartphone. That said, internet access is cost-prohibitive for most community members given their economic status.

Without proper internet access, these communities will continue to struggle to bring their languages online, resulting in a circular dynamic regarding the lack of resources in critically endangered languages.

3.5. Documentation and study

When critically endangered languages face limited language documentation and studies, the perspectives for Languages Technologies are tortuous to say the least. Is not a coincidence that HRL and MRL are among the best documented and most studied languages, while a good amount of LRL are currently under-documented and under-studied like Ayapaneco.

When documentation and studies of a language increase, so do the chances of implementing Language Technologies. Consequently, it is fundamental to improve the documentation and study of critically endangered languages, especially those that are among the least documented and studied.

3.6. Linguistics

Language variation can pose a challenge for Language Technologies in critically endangered languages. Variation is an intrinsic characteristic of human language, and it follows the orderly heterogeneity premise (Weinreich et al., 1968) under which language variation can be conditioned by either linguistic or social factors, or the interaction of both. While variation is widely attested in all world languages, the study of critically endangered languages has recently called into question the orderly heterogeneity premise. Indeed, critically endangered languages exhibit a large proportion of unstructured variation that cannot be linked to social or linguistic factors (Dorian, 2010). My recent research on Ayapaneco shows that this language also exhibits a high proportion of unstructured variation, thus confirming the trend found in other critically endangered languages (Rangel, 2019).

Unstructured variation has been left out of Language Technologies. This is understandable considering that until recently, it has also been overlooked by the fields traditionally concerned with variation in general. Furthermore, as unstructured variation is widely present in critically endangered languages that also happen to be LRL, this contributes to the existing blind spot in modeling, processing and analyzing this type of variation and hinders the documentation, description, and revitalization of these languages as well as the implementation of Language Technologies.

4. Conclusion

The implementation of Language Technologies in critically endangered languages poses complex and multifaceted challenges such as the annotation bottleneck, heavy limitations in human resources and capacity, scarce infrastructure, limited documentation and study, as well as under-studied linguistic particularities.

These multifaceted challenges seriously hinder future perspectives not only for Language Technologies but also for the documentation, preservation, reclamation, and revitalization of critically endangered languages. Consequently, it is imperative to think outside of the box to apply these technologies as they could help maximize the limited

time we have left to engage with critically endangered languages. As these languages represent almost 10% of the world's total, and could disappear at any time in the next decade, time is of the essence and this task should be prioritized.

It is no longer enough to have just a few isolated experts and community members working to document and study these languages as the annotation bottleneck severely limits language reclamation and revitalization efforts as well as the deployment of Language Technologies. On the contrary, concrete and multifaceted measures must be taken immediately involving new multidisciplinary approaches while creating synergies among varied actors (academia, governments, NGOs, communities and civil society) to better address these challenges and support capacity building in the long term for these communities. Simultaneously, and more fundamentally, we must address the root causes of these challenges such as inequality, poverty and discrimination as is not a coincidence that critically endangered languages are also low-resourced. Thus, the lack of resources and Language Technologies replicates those social, political, and economic inequalities existing in the world. The next few years will be decisive in attempting to break this circle before a significant proportion of languages disappear from the face of the Earth.

5. Bibliographical References

Ćavar, M., Ćavar, D., and Cruz, H. (2016). Endangered language documentation: Bootstrapping a chatino speech corpus, forced aligner, ASR. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4004–4011, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Dorian, N. (2010). *Investigating variation: the effects of social organization and social setting*. Oxford studies in sociolinguistics. Oxford University Press, Oxford-New York.

Duong, L. (2017). *Natural Language Processing for Resource-Poor Languages*. PhD dissertation, University of Melbourne.

Henrich, J., Heine, S. J., and Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3):61–83.

King, B. P. (2015). *Practical Natural Language Processing for Low-Resource Languages*. PhD dissertation, University of Michigan.

Moseley, C. (2010). UNESCO Atlas of the World's Languages in Danger. Online version: `ttp://www.unesco.org/culture/en/endangeredlanguages/atlas`.

Rangel, J. (2017). Les derniers locuteurs : au croisement des typologies des locuteurs de langues en danger. *Histoire Epistémologie Langage*, 39(1):107–133.

Rangel, J. (2019). *Variations linguistiques et langue en danger. Le cas du numte oote ou zoque ayapaneco dans l'état de Tabasco, Mexique.* Thèse de doctorat, INALCO.

Seifart, F., Evans, N., Hammarström, H., and Levinson, S. (2018). Language documentation twenty-five years on. *Language, Journal of the Linguistic Society of America*, 94(4).

Weinreich, U., Labov, W., and Herzog, M. (1968). *Empirical foundations for a theory of language change*. University of Texas Press, Austin, Texas.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). ELAN: a professional framework for multimodality research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May. European Language Resources Association (ELRA).

Proceedings, pages 324–327
Paris, UNESCO Headquarters, 5-6 December, 2019.

Situation and Challenges of Technologies for Indigenous Languages of India

S Sinha, S S Agrawal
Amity University Haryana, India[1], KIIT College of Engineering, Gurgaon, India[2]
ssinha@ggn.amity.edu[1], ss_agrawal@hotmail.com[2]

Abstract

India is a country with huge linguistic diversity. Out of 900 languages spoken in the country, only a few have witnessed the digital world. This paper presents the language situation in India. It also highlights the opportunities, barriers and complexities faced by the language technology community in the development of indigenous Indian languages. The aim is to study their influence on the adoption and adaptation of digital technology. Technological achievements/fallouts of Indian languages relating to the world languages will be analysed with the purpose to identify the gap. The paper also outlines the need for future language resources and uptake of projects for technological advancements of indigenous languages of India.

Keywords: Language Technology, Indian languages, Under-resourced Language

Résumé

भारत एक विशाल भाषाई विविधता वाला देश है। देश में बोली जाने वाली 900 भाषाओं में से कुछ ही डिजिटल दुनिया में देखी गई हैं। यह पेपर संसाधनों, और प्रौद्योगिकियों के संदर्भ में भारतीय भाषाओं की स्थिति को विस्तार से प्रस्तुत करता है। यह भारतीय भाषाओं की प्रौद्योगिकियों की विशिष्ट आवश्यकताओं, अवसरों, बाधाओं और जटिलताओं को उजागर करता है। इसका उद्देश्य डिजिटल प्रौद्योगिकी को अपनाना और उनके अनुकूलन पर उनके प्रभाव का अध्ययन करना है। विश्व भाषाओं से संबंधित भारतीय भाषाओं की तकनीकी उपलब्धियांऔर अंतर की पहचान कर भविष्य की परियोजनाओं को पूरा करने की आवश्यकता है।

1. Introduction

India is a plurilingual and pluri-ethnic land. Linguistic diversity and multilingualism are essential for the enrichment of humanity and development and language is an important attribute of it. According to the census (2011), there are 121 languages and around 2300 dialects in India. These languages belong to five language family(census, 2011): the Indo-European (Indo-Aryan 78.05%), Dravidian (19.64%), Austro-Asiatic, Tibeto-Burmese and Semito-Hamitic. Out of all the languages spoken in the country, 22 languages are constitutionally recognized and 'Hindi' has the status of the official and national language (Jha, 2010). Table 1 presents the language and speaker population of major languages of the country.

In India, there are around 30 languages with more than one million population, but most of them have not seen the light of the digital world. This situation puts the users of indigenous languages in a disadvantageous situation. It creates a digital divide among the languages and puts them in danger of digital extinction that may lead to complete extinction also. The development of language technologies provides opportunities to exchange ideas with one another easily. Research community working in the area of language technology look forward to utilizing technological growth to create a workable platform. They aim to cater to the need of users irrespective to their language, age, gender and socio-economic background.

Undoubtedly, long term effort is required to cover all the languages and take benefit from digital growth. This paper highlights the language situation and technological growth for Indian languages. The analysis of the current situation helps to identify the existing challenges and barriers for language users. In the end, the paper outlines the need for future language resources and uptake of projects for technological advancements of indigenous languages of India.

Languages	Population(%)	Languages	Population(%)
Hindi	43.63	Malayalam	2.88
Bengali	8.03	Punjabi	2.74
Marathi	6.86	Assamese	1.26
Telugu	6.70	Maithili	1.12
Tamil	5.70	Santhali	0.61
Gujrati	4.58	Kashmere	0.56
Urdu	4.19	Nepali	0.24
Kannada	3.16	Sindhi	0.23
Odia	3.10	Dogri	0.21

Table 1: Major languages and speaker population of India

2. Digital Representation of Indian Languages

The three Indian Languages; Hindi, Punjabi and Bangla are among the top ten most widely spoken languages of the world (Arora et al.,2013), but none of these finds their place in the top ten languages on the web (Sinha et al.,2018). According to Unesco's "Atlas of the world's languages in danger " (Language atlas, 2009). India has the maximum number of endangered languages and most of the Indian languages are vulnerable. Availability of digital data for the languages may help in their revival. Online services in these languages may increase their user base. Limited language support and content are the largest barriers to the adoption of online services.

The internet contents are majorly available in the languages of developed countries with English having the topmost share with 56%, followed by Russian and Spanish with

7.3% and 4.7% respectively (Arora et al., 2013). Only 0.1% of the web content is available in Hindi language and none of the other Indian languages finds its place in the top 40 languages of the digital world. As far as language technology is concerned Hindi, Bangla, Telugu and Tamil are a few Indian languages that have some associated language technology with various quality level. Lack of digital resources for the Indian languages categorize them as under-resourced languages. Technology development is essentially required for keeping these languages alive. Efforts have been made to provide technical support to the Indian languages, but lack of resources makes it a challenging task. To start with, the researchers have worked for the technology development for few languages as mentioned above and is in the process of generating resources for many more languages.

3. Technology Development for Indian Languages

Resource creation is the first and the foremost important step towards the technology development for languages. Indian languages falling in the category of under-resourced languages require special effort for corpus creation. Efforts made in this direction have helped the researchers to create corpora for a few Indian languages and application based on ASR, TTS and MT have been developed for a few languages.

3.1 Development of Language Resources

Language technology is a data-centric research area. Text and speech data of any language are the necessity for developing technology for that language. Several research groups, the Ministry of Human Resource Development (MHRD) and its agency for language development; CIIL (Central Institute of Indian Languages) and Ministry of Electronics & Information Technology (MeitY) with its agency TDIL (Technology Development for Indian Languages) is continuously putting efforts for developing Indian languages resources.

3.1.1 Text Corpus Availability

Collection of phonetically rich text sentences are available for Malayalam, Kannada, Marathi, Hindi, Tamil, Punjabi, Bangla, Indian English and Assamese. The size of these corpus ranges from 3000 sentences to around 23000 sentences. Only a few of these databases are multilingual, and most of it is monolingual.

3.1.2 Speech Corpus for Indian Languages

The development of speech corpus is a time and labour intensive task. The created corpus has to be annotated before being utilized for technology development. The corpus collection was initially done in the studio environment with the aim to reduce noise and external interferences. Studio recording posses restriction in mobility. Since most of the Indian languages are under-resourced so availability of speakers for studio recording is difficult. With the advancements of technology, availability of noise reduction techniques and to match with the real-time scenario, studio recordings are not being used nowadays. The demand for corpus created in an office environment or noisy areas like roadside, moving vehicle or market place prevails now. Efforts are continued to collect resources using crowdsourcing (Arora et al., 2016) or through online platforms (Sinha et al.,2017) as a read or spontaneous speech. Table 2 presents the statistics of speech resources developed and reported so far.

Resources	Language & Statistics	Organization
PLS	Hindi: 50,000 lexemes, Marathi: 51,065 lexemes, Punjabi: 33,874 lexemes, Manipuri: 2,83,998 lexemes Assamese: 53,304 lexemes	TDIL
Speech samples: agriculture domain	Telugu 1073 speakers, Tamil 1000 speakers, Marathi 1500 speakers, Bangla 1000 speakers, Assamese 1023 speakers	TDIL
Annotated speech samples	Bengali 450 speakers, Hindi 650 speakers, Konkani 450 speakers, Odia 450 speakers, Malayalam & Tamil 450 speakers	LDC-IL
Global Phone	2000 native speakers transcribed data in Tamil	ELRA
EMILLE/ CIIL Corpus	Monolingual, parallel and annotated corpora in Assamese, Bengali, Gujarati, Hindi, Kannada, Kashmiri, Malayalam	ELRA
Annotated Speech Samples	Assamese: 5658 speech data files; 27 speakers, Bengali :2500 speech data files; 21 speakers, Nepali: 660 speech files; 6 speakers, English :2500 speech files; 16 speakers	IIT Guwahati
Prosody model develop-ment	Gujrati:1000 speakers IVR recording, Audio search system, ASR Marathi: 1000 speakers IVR recording, Audio search system, ASR	DAICT, Gandhi Nagar
Prosodic word Dictionary	English: 5031-word dictionary generated from 2500 spoken Bengali sentences	IIT Kharagpur

Table 2: Speech resources available for Indian languages

3.2 Development of Language Technology

Every human being wishes to deliver and also obtain information and services in their language. Today, for linguistic preservation and cultural redemption, development of language technology and digital representation of languages has become essential. Application based on automatic speech recognition, text to speech synthesis and machine translation makes life easier for people who like to avail facilities in their native language. Some efforts in this direction have been made for a limited number of Indian languages.

3.2.1 Automatic Speech Recognition (ASR)

Literature (Singh et al.,2019) highlights that the language research community of India has carried out several experiments for different Indian Languages based on small databases collected for experimental purpose. Most of these are done as a laboratory experiment and are not

converted into applications for general use. Researchers and industry have majorly focussed on HMM and ANN based methodologies for the development of ASR systems. A big giant like Google has created Assistant that converse in 6 Indian languages apart from many world languages. But, again this is a very small fraction as compared to 121 languages. Several prototypes such as railway enquiry system (Samudravijaya,2000), Bangla digit recognizer, travel enquiry system etc. have been developed by Indian research institutes, A major breakthrough in this direction is achieved by the development of a system for agriculture commodity prices. Speech-based access for Agricultural Commodity prices for 6 Indian Languages was developed for Hindi, Bengali, Assamese, Tamil, Telugu and Marathi. The project was carried as a consortium project supported by DeitY, India. The system uses an HMM-based speech recognizer and is helpful to illiterate farmers and visually impaired people. But, again this type of system lacks in catering to the dialectal, prosodic and tonal variations.

3.2.2 Text to Speech Synthesis (T-T-S)

TTS system when integrated with a screen reader is potentially assistive technology for visually impaired people. Concatenative and statistical approaches have been used to develop TTS engine for some of the Indian languages. TTS applications developed so far for Indian languages are as mentioned below(TDIL):

- **TTS integrated with Screen Reader for Visually Challenged persons:** TTS integrated with Screen Reader are available in Hindi, Bengali, Marathi, Tamil, Telugu and Malayalam.
- **Browser Plug-in**: TTS as browser plug-ins are also developed for eight Indian Languages namely Hindi, Bengali, Marathi, Tamil, Telugu, Malayalam, Odia and Gujarati.
- **SMS Reader in Indian Languages - Sandesh Pathak**: SMS Reader is an Android App and is made available for 5 Indian Languages namely Hindi, Marathi, Tamil, Telugu and Gujarati. Click here to download

These projects were carried out in consortium mode under the leadership of IIT Madras. Apart from this few works in this direction for other languages have also been reported. Table 3 presents the details.

Sl No	Name of the Language	Development of TTS Engines (Concatenative and Statistical approach)
2.	Assamese	Male and Female voice – HTS, USS
3.	Bengali	Male and Female – HTS, USS
4.	Gujarati	Male HTS, USS, Male HTS using STRAIGHT approach.
5.	Marathi	Male and Female – HTS, USS, Male HTS STRAIGHT
6.	Malayalam	Male and Female – HTS, USS
7.	Kannada	Male HTS

Table 3: Details of TTS system development efforts

3.2.3 Machine Translation

In a multilingual country like India, a huge amount of information exchange takes place across different languages inside the country. It is thus necessary to have an automated process to convert data from the sender's language to that of the receiver's language. Efforts have been put in this direction to obtain automatic translator of languages. Most of the work till date is confined between Hindi and English language pair but other language pair are also being tried now. Table 4 presents some of the well known MT systems available in Indian languages. The approach used by them ranges from example-based to rule-based and to statistical MT systems.

System	Target Language	Place	Features
Mantra	English to Hindi, Gujrati, Telegu. Hindi to English, Bengali, Marathi	CDAC,Pune	Uses Tree Adjoining Grammar Formalism.
Anubaad	English to Bengali	CDAC, Kolkata	A hybrid system which uses n-gram approach for POS tagging. Works at sentence level
Anglabharti, AnglaHindi, Anubharti	English to Hindi, Tamil	IIT,Kanpur	Uses intermediate structure Pseudo Lingua for IL.
English Hindi MTS	English to Hindi	IIIT, Hyderabad	Combines Rule Based Machine Translation and phrase based SMT

Table 4: Machine translation systems in Indian languages

Speech to speech translation system has also been tried upon. CDAC Kolkata developed a prototype for Hindi-Bangla speech to speech dialogue system. A consortium project for speech to speech translation system was initiated at the international level and India was also a part of it (Arora et al.,2013).

4. Challenges in Technology Development for IL

In the race of language technology, Indian languages lie far behind the languages of other developed countries. The major requirement is for resource creation based on global standards. Apart from this, several other issues influence the technology growth for languages especially that of under resource languages. Some of the challenges faced by Indian languages for technology development is as follows :

- Language ambiguity and complexity: the same word has a different meaning when used in a different context
- Origin of Indian script and family: One language is represented using many scripts and also many languages follow the same script.
- Difficulty in data collection due to geographical, social and cultural strata of the country.

- Presence of several dialects: code-mixing between dialects; a massive number of non-native speakers of languages.
- Non-conformance with English centric models: existing models can't be extended to Indian languages.
- Localization issues associated with the operating system, keyboards and applications.
- Lack of encoding standards: several phones of Indian languages have not yet been encoded in existing standards

5. Way Forward for Indian Language Technology Development

Rigorous efforts are required to bring Indian languages on the world map of technologically developed languages. Some of the tasks are identified below to help curb the challenges faced in technology development for these languages:

- Producing a White paper: Necessary to reflect the current situation for all languages.
- Massive amount of text data creation to reliably train a statistical language model: focus should be on phonetically balanced data.
- Obtain transcribed recordings from several speakers to capture varying acoustical characteristics due to nativity and other sociolinguistics aspects for creation of the acoustic model.
- Pronunciation dictionary of the vocabulary for lexical/PLS development: the focus should be on capturing prosody.
- The urgency to work with zero resource language: use an AI approach and try to avoid its extinction.
- Generate facilities such as BLARK (basic language resource tool) for all IL.

6. Conclusion

Language technology contributes to promoting linguistic diversity and multilingualism in the digital world. Now, the technology is moving into the daily life of people in the different application area. India is a country with diverse linguistic variations. Very few Indian languages have been worked upon for the development of language technology. The present paper highlights the technological achievements of Indian languages. Many languages have shown their presence in the digital world and efforts in this direction is still continued. But, to date, the indigenous people are still experiencing barriers to access information through the internet. They experience obstacle to use a tool not available in native languages. Indian languages being under-resourced face more difficulty in this regard and may require a long term effort to get benefit from the latest digital developments.

7. Bibliographical References

Arora S., K. K. Arora, M. K. Roy, S. S. Agrawal, and B. Murthy, (2016). Collaborative speech data acquisition for under resourced languages through crowdsourcing," Procedia Computer Science, vol. 81, pp.37-44.

Arora, K., Arora, S. & Roy, M.K. (2013) Speech to speech translation: a communication boon. *CSIT* 1, 207–213 (2013) doi:10.1007/s40012-013-0014-4

Jha, G. N. (2010). India's language diversity and resources of the future: Challenges and opportunities. *Special Center for Sanskrit Studies, Jawaharlal Nehru University, New Delhi*.

Samudravijaya, K.(2000). Computer Recognition of Spoken Hindi‖. Proceeding of International Conference of Speech, Music and Allied Signal Processing, Triruvananthapuram, pages 8-13, 2000.

Singh, A., Kadyan, V., Kumar, M. *et al.(2019).* ASRoIL: a comprehensive survey for automatic speech recognition of Indian languages. *Artif Intell Rev* (2019) doi:10.1007/s10462-019-09775-8

S Sinha, S Sharan, S S Agrawal,(2017). O-MARC: A multilingual online speech data acquisition for Indian languages, Oriental-COCOSDA , Nov 1-3, 2017, held at Seoul, S Korea.

Sinha Shweta, Shyam S Agrawal (2018). Sustaining Linguistic Diversity Through Human Language Technology : A Case. Study for Hindi.May 2018. CCRUL-LREC 2018

Government of India, <www.censusindia.gov.in/ 2011Census/C-16_25062018_NEW.pdf> Accessed 08/01/2020

Governemnt of India, <tdil-.in/index.php?option= com_vertical & parentid =85 &lang=en> Accessed : 12/12/2019

Unesco, 2009 <www.unesco.org/languages-atlas/ index.php> Accessed 24/12/19

Proceedings, pages 328–331
Paris, UNESCO Headquarters, 5-6 December, 2019.

Towards ASR that Supports Linguistic Diversity in Norway

Benedicte Haraldstad Frostad, Verena Schall, and Sonja Myhre Holten
The Language Council of Norway
Observatoriegata 1 B, 0254 Oslo, Norway
{benedicte.frostad, verena.schall, sonja.myhre.holten}@sprakradet.no

Abstract

Norway's majority language, Norwegian, has two written standards, many dialects and no spoken standard. Norway also officially recognises some national minority languages. The extra costs, need for linguistic expertise and lack of suitable lexica and speech data sets complicate the development of ASR products for all these Norwegian language communities. This poses a democratic problem as public institutions automatise dictation and integrate ASR as a means for interaction. the Language Council is initiating innovative projects to improve ASR for Norwegian and minority languages in Norway and wishes to exchange ideas and experiences.

Keywords: linguistic diversity, dialects, written standards, spoken standards, automatic speech recognition, Norwegian, minority language, Norwegian sign language, indigenous languages, Kven, Sami, Romanes, Romani

Résumé

Majoritetsspråket i Norge, norsk, har to offisielle skriftnormer, mange dialekter og ingen offisiell uttalenorm. Norge annerkjenner også noen nasjonale minoritetsspråk. De ekstra kostnadene, behovet for lingvistisk ekspertise og mangelen på passende leksika og taledatasett vanskeliggjør utvikling av talegjenkjenningsprodukter for alle disse språksamfunnene. Det er et demokratisk problem ettersom offentlige institusjoner innfører automatiske dikteringsverktøy og integrerer talegjenkjenning i sine kommunikasjonskanaler. Språkrådet har tatt initiativet til nyskapende prosjekter skreddersydd for å øke kvaliteten på talegjenkjenning for norsk og minoritetsspråk i Norge og ønsker å utveksle idéer og erfaringer.

1. Introduction

The lack of support for one's mother tongue in services and products with integrated automatic speech recognition (ASR) represents a challenge to European and national aims to ensure equal participation in society for all citizens (De Smedt 2012: 41, Directorate-General of the UNESCO 2007). The situation is pressing in Norway, where public and private institutions are increasing integration of ASR. Notably, the courts and The Storting, the Norwegian parliament, are initiating automatic dictation of all court and parliament sessions. This is a challenge in the majority language, Norwegian, which has diversity in both written and spoken forms that is unusual for a national language, and an even greater challenge for minority languages.

Norway has a long history of Norwegianisation policy directed towards sign language users, indigenous and national minorities. This policy has been carried out differently towards each community, thus, the vitality of Norwegian minority languages and their language communities varies, as do the resources available for each language. This is particularly noticeable in the field of language technology, where research and development rely heavily on language resources of high quality and linguistic expertise.

2. Linguistic diversity in Norway

2.1 Norwegian

Norwegian is a North Germanic language with approximately 5 mill. speakers, closely related to Swedish and Danish, and it is the majority language in Norway. It has two written and no spoken standard. It has a large number of dialects with significant phonetic, lexical and syntactic variation. (Skjekkeland 1997). There is no spoken variant with a status as an official language.

2.1.1 Written Norwegian

None of the two written standards for Norwegian, Nynorsk (NN) and Bokmål (NB), can be said to correspond to a certain spoken variety. *Målloven* (the Language Act) regulates the use of written standards in the public sector, where each must be used in min. 25% of all text. Both NN and NB allows for significant lexical and inflectional variation. De Smedt and Rosén (1999) demonstrates how a long sentence in Bokmål may be spelled in no less than 165,888 different ways. There is also extensive code-switching between the two.

2.1.2 Spoken Norwegian

Norwegian dialects have a more prominent role than in other European countries, due to the lack of an official and even a de facto spoken standard (De Smedt et al. 2012: 45), and dialectal variety makes use of ASR challenging in Norway.

2.2 Norwegian Sign Language

Norwegian Sign Language (NSL) is among the largest minority languages in the country, with an estimated 16,500 speakers (the Norwegian Ministry of Culture 2008). Its status is nonetheless precarious. During the last decade three out of four national deaf schools have closed down, as a new policy to integrate children with hearing loss in mainstream schools was introduced. Negative attitudes to sign language and the loss of communities due to inclusion in mainstream schools, have led to a less vital status. Political aims for NSL are two-fold. The first is the so called democratic perspective (as stipulated in the Convention on the Rights of Persons with Disabilities (CRPD)) where key words are, access (to the mainstream society), and facilitation. Access and facilitation, because signers who are deaf or hard of hearing are more or less relying on sign language, as this is the only language

modality in which they can interact freely and effortlessly with their interlocutors. The second is the language policy perspective. NSL has a value in itself. It is a culture and an identity marker, as much for hearing people as deaf/hard of hearing, and in addition part of our cultural heritage. Today, the NSL operates with two geographical variants, one in the north part of the country, and one in the south.

NSL has been legally recognised in Norway through the Education Act as a minority language since 1997. Norway have also ratified CRPD, and has thus committed to promote and recognise the use of sign language, and to implement universal design. NSL was recognised as a part of the Norwegian language diversity and a part of the cultural heritage of Norway in The Norwegian Ministry of Culture (2008) and this right was affirmed in the Storting in 2009.

2.3 The Sami languages

The Sami languages belong to the Uralic language family. There are mainly three Sami languages in use in Norway: North, Lule, and South Sami, which to some extent are mutually intelligible. The North Sami Language has the largest language community. In comparison, Lule Sami and South Sami are far more endangered with fewer language users and fewer resources.

2.4 Kven

The Kven language belongs to the Finno-Ugric group of the Uralic language family. It is heavily influenced by Norwegian and the Saami languages, and closely related to Meänkieli, a national minority language in Sweden. Due to suppressive national language policies, the Kven ethnic minority outnumbers the language community. There are no official numbers, but one presumes that the former counts about 50,000-60,000 members, whereas estimates for the number of speakers range between 1500 and 10,000 (Schall 2017).

2.5 Romani

Romani (also known as Scandoromani) is spoken by the Romani ethnic minorities (some members of the community prefers to be referred to as Travellers) in Sweden and Norway. It has a North Germanic grammatical structure and a heavy lexical influence by Romanes.

2.6 Romanes

Romanes is spoken by the Roma ethnic minority in Norway, and is internationally often referred to as Romani. It is an Indo-Aryan language with many varieties all over Europe. The Norwegian varieties belong to the Northern Vlax-group. Bilingualism between several variants is fairly common, and many speakers speak two or more variants, , especially Lovara and Kalderaš. It is estimated that the Roma community in Norway count approximately 700 individuals, and the continued marginalisation of the community has a significant impact on the language community. Studies indicate a lack in children's formal education and a high illiteracy rate among adults (Hagatun 2019). Romanes is mainly used as a spoken language, and there are few written sources of the language in Norway. Most Roma children start school mainly Romanes speaking, an indication of a healthy spoken language and conscious language planning in families and within the community. Unfortunately, authorities have shown little initiative to protect the language. Developing language infrastructure is essential and will be valuable for educational practice, but as the varieties of Romanes spoken in Norway are still poorly documented, this is a challenging task.

2.7 Yiddish

Yiddish is a West Germanic language with considerable influence from Hebrew, Aramaic and Slavonic languages, with a long history as a minority language in Norway. It is not recognised in the European Charter for Regional or Minority Languages (ECRML) The Language community, however, has certain language-related rights through the Framework Convention for the Protection of National Minorities, but unfortunately this framework is less specific than ECRML. Yiddish was a vital language pre-WWII, but is now near-extinct in Norway.

3. Language policy ambitions

Norwegian language policy aims to ensure that everyone has the right to a language, to evolve and acquire the majority language, Norwegian, and to evolve, acquire and use their mother tongue, including Sign Language, indigenous languages or national minority languages (the Norwegian Ministry of Culture 2008: 24). To meet these goals, it is important that language use in digital interactions are not left out of the picture.

3.1 Language policy and ASR

Norwegian courts and the parliament are initiating fully automated dictation for parliament and court sessions. Furthermore, an increasing amount of private and public institutions are communicating with users by means of chatbots, and expect to integrate ASR in these services for increased streamlining and as a means to enable universal design. the Language Council of Norway is responsible for informing institutions about the challenges involved with ASR development and the support of linguistic diversity, as well as to work towards better enabling institutions making use of ASR, and to better enable developers and researchers to provide products suited for the various Norwegian language communities.

3.1.1 The Norwegian Language Bank

Following up on the language policy ambitions stipulated in the parliament white paper *Report no. 35 (2007-2008) to the Storting* (The Norwegian Ministry of Culture 2008), the Norwegian parliament made funds available for a national language bank in 2010, with the aims to collect resources for use in language technology research and development, such as large datasets for text and spoken language, and lexica, available to public as well as private institutions. The National Library is currently responsible for hosting the Language Bank, where resources can be downloaded with no registration necessary by anyone. In 2019, the Parliament decided to make funds available for the development of new resources and it has been decided that The National Library and the Language Council of Norway plan which resources should be developed and made available in co-operation.

4. Linguistic diversity and ASR

4.1 Challenges

4.1.1 Scarcity of data

Training an acoustic model and language model for the development of ASR requires sufficient annotated speech data, a pronunciation lexicon (in most cases) and sufficient text data.

The availability of language technology support for Norwegian is extensive considering the size of the language community. However, the lack of a spoken standard and the two written standards makes most products unavailable to a considerable number of speakers. With a few exceptions, products only support one written standard, Bokmål, and speech technology products only support the dialect spoken in the region of the capital, Oslo. Supporting the linguistically diverse Norwegian language community requires more linguistic resources, tailor-made to address linguistic diversity by teams including linguists with expert knowledge of written and spoken or signed variants of Norwegian and Norwegian minority languages.

4.1.2 Costs

Supporting the degree of language variety that is necessary for the development products and services with integrated ASR that can be used by all Norwegian and minority language speakers is costly. Support for Norwegian requires more resources than for languages such as Dutch and Swedish, where speakers can make use of spoken standards. Norwegian speakers who are not recognised by ASR software in their regional dialect, have no means to standardise their language in order to be understood. There are few resources available as of yet to support the minority languages. It is therefore important that the government provide resources of high quality, for use by developers and researchers.

4.1.3 Linguistic expertise

The Language Council has learnt through interviews with the developers that computational linguists with sufficient proficiency in the Nynorsk written variant, as well as Norwegian spoken dialects are hard to come by, particularly for developers based outside Norway where most development of Norwegian ASR takes place. Linguistic expertise in the minority languages is significantly scarcer.

4.1.4 Data sharing and information exchange

To ensure an efficient use of resources and funding, it is vital that information on products, available resources and linguistic expertise is shared between language communities, developers, institutions making use of products and services with integrated ASR, researchers and the Language Council. There is currently no efficient infrastructure for the exchange of such information, and the establishment of good networks is a necessary first step towards speech technology that supports Norway's linguistic diversity, and meet language-related political aims.

4.2 Existing resources

4.2.1 Resources in the Language Bank

The Language Bank contains several resources for language technology in Norwegian, developed as part of public research projects or by private companies. As of now, it contains few resources suited to address spoken and written diversity support needs in ASR development and no relevant resources for minority languages.

4.2.2 Resources in CLARINO

Resources available in the Norwegian CLARIN database are mostly suited for research in language technology and linguistics. CLARINO data mostly covers Norwegian, but will contain a large dataset for NSL in the near future.

4.2.3 Resources in Giellatekno and other institutions

Resources for the Saami languages and Kven are largely managed in cooperation by Divvun, the Kven Institute and Giellatekno, an open source repository. Some text processing tools and text databases exist for Kven. There are a few more for the Saami languages, due to a TTS project There are pronunciation lexica, text and speech databases and various other tools for Sami languages. There are some lexical resources for Norwegian Romani made available on https://app.uio.no/hf/nro/index.php?link=contact from a Ph. D. project.

4.3 Current initiatives

The Language Council has initiated two new resources to be developed in co-operation with the Language Bank in 2020 to enable support for dialect diversity and both written standards in Norwegian in ASR development. One is an extension of an already existing pronunciation lexicon with additional transcriptions representing the pronunciation of lexicon items in four additional dialects. The dialect variant spoken in the Oslo region is already represented in the lexicon. The dialects are carefully selected in co-operation with the University of Oslo, to represent all five dialect areas, and as much lexical and phonetical variation as possible. All transcriptions will be tagged for dialect, and developers and researchers can select the ones they want to include. The pronunciation lexicon can be used for both ASR and speech synthesis (TTS). The other is a speech database covering all five aforementioned dialects, tailor-made for digital assistants, notably automotive and mobile assistants. The National Library has taken the initiative to create a large speech database consisting of annotated parliament sessions. A wide variety of dialects are used in parliament, and the annotation will be available in both written forms. the Language Council is reaching out to stakeholders to plan projects aimed at ASR support for Norwegian minority languages.

NSL is a visual-gestural language and like all signed languagesit lacks a written form and is primarily a face-to-face language. Thus, in order to create accessible and open documentation of NSL, a digital (online) platform able to host a large collection of video-recordings of the language is needed. Current initiatives in Norway, such as the CLARINO infrastructure project, show potential, but financial and technological support is needed to ensure that documentation of NSL can be archived and preserved over the long-term.

4.4 Towards ASR that supports linguistic diversity

4.4.1 Close co-operation with developers, researchers and language communities

To ensure that funding for the development of resources to be used for research and development of language technology is used in an as efficient way as possible, the Language Council needs close communications with resources and institutions planning on using language technology resources. It is important to put emphasis on products that are under development or will be developed and used in the near future, rather than products that may not be developed anytime soon. Thanks to efficient communication with developers and public institutions, the Language Council learned that it is currently vital to create speech datasets for automotive and mobile assistants and parliament sessions. However, we can do better. An efficient infrastructure for information exchange must be established to provide the resources and advise needed to researchers, developers, public institutions and, above all, the users, the members of the many Norwegian speech communities themselves, that need access to new technology.

4.4.2 Sign language recognition

A project needs to be initiated and funded where gesture recognition researchers and developers collaborate with NSL linguists on exploring the possibility for NSL recognition.

4.4.3 ASR development for under-resourced languages

The Language Council and developers are collaborating on exploring the possibilities to develop ASR for the Kven language, which could possibly be combined with the closely related language Meänkieli, spoken in Sweden. Based on comparable projects, where e.g. ASR for Afrikaans was successfully developed with a system partly trained on Dutch, the possibility of using an acoustic model based on Finnish, a closely related language, is explored. Unfortunately, the Sami languages do not have well-resourced languages that are close enough in linguistic proximity for this to be an opportunity.

4.4.4 Universal design

It is vital that universal design is considered at all times when planning and developing new resources. Norway's Equality and Anti-Discrimination Act stipulates that all public and private undertakings focused on the general public have a duty to ensure that their general functions have a universal design, including all ICT solutions. Considering the limited availability of funds, developers and linguistic expertise, it is important that all resources that are developed are well planned and contribute to the development of technology that satisfies legal requirements.

4.4.5 Development and dissemination of linguistic resources

Norwegian and all minority languages in Norway are severely under-resourced. As of today, no pronunciation lexicons exist for any minority language, and there is only one for 350,000 words in Norwegian Nynorsk. Speech databases for Norwegian lack sufficient dialectal coverage and coverage for age and gender. the Language Council is reaching out to all stakeholders to plan the development of quality-assured resources suited to develop ASR technology that works for all. Some projects are under way, but many more are needed.

5. Conclusion

More funding and governmental initiative are needed to secure ASR that is accessible to speakers of Norwegian and Norwegian minority languages. Long-term planning and stable, large projects are needed to provide resources of the quality that is needed for ASR and sign language recognition that supports Norway's linguistic diversity. Careful planning of resources requires an infrastructure that allows for efficient information and data exchange between representatives of the language communities, the Language Council, research communities, developers, public institutions, and institutions developing resources and making them available (through interfaces such as the Language Bank). It is particularly beneficial for sign language recognition and ASR development for minority languages that this infrastructure also includes international collaborators. Governmental institutions need to make informed decisions when applying ASR, and take linguistic diversity into consideration. the Language Council has addressed this issue when reaching out to and meeting relevant institutions, and this needs to be followed up.

6. Bibliographical References

De Smedt, Koenraad, Gunn Inger Lyse, Anje Müller Gjesdal and Gyri S. Losnegaard. 2012. *The Norwegian Language in the Digital Age – Norsk i den digitale tidalderen*, METANET Whitepaper, Berlin: Springer.

De Smedt, Koenraad and Victoria Rosén. 1999. «Automatic proofreading for Norwegian: The challenges of lexical and grammatical variation» in *Proceedings of NOVALIDA 1999*.

Directorate-General of the UNESCO. 2007. *Intersectional Mid-term Strategy on Languages and Multilingualism*, http://unesdoc.unesco.org/images/0015/001503/150335e.pdf

Hagatun, K. 2019. "They assume that I don't really want education for my children": Roma mothers' experiences with the Norwegian educational system" in *HERJ Hungarian Educational Research Journal. Special issue., 9*(1), 9-21. doi:10.1556/063.9.2019.1.2

The Norwegian Ministry of Culture. 2008. *Report no. 35 to the Storting (2007-2008): Mål og Meining – Ein heilskapleg norsk språkpolitikk*, Oslo: Akademika AS.

Schall, V. 2017. «Språk, identitet og minoritetspolitikk [Language, identity and minority policy]» in N. Brandal, C.A. Døving, & I. Thorson Plesner (Eds.): *Nasjonale minoriteter og urfolk i norsk politikk fra 1900 til 2016*. Oslo: Cappelen Damm Akademisk.

Skjekkeland, Martin. 1997. *Dei norske dialektane – Tradisjonelle særdrag i jamføring med skriftmåla*, Kristiansand: Høyskoleforlaget.

Proceedings, pages 332–334
Paris, UNESCO Headquarters, 5-6 December, 2019.

European Language Monitor – Exploring European Language Policies On-Line

Sabine Kirchmeier
European Federation of Official Institutions for Language (EFNIL)/Danish Language Council
Adelgade 119 B, DK-5400 Bogense
sabine@dsn.dk

Abstract

The European Language Monitor is a project of EFNIL, the European Federation of National Institutions for Language. Since 2009, the project has collected and published data on language legislation and language planning in Europe. This article gives an overview of the information that is provided on European language legislation and language planning based on an extensive questionnaire with 9 sets of questions. Data for more than 25 European countries is accessible via a web interface. The paper argues that access to information on language policies across countries is an important asset for policy makers, journalists, scientists and teachers.

Keywords: language policy, language strategy, data collection.

Resumé

European Language Monitor er et projekt under EFNIL, den europæiske sammenslutning af nationale sproginstitutioner i Europa. Siden 2009 har projektet indsamlet og publiceret data vedr. sproglovgivning, sprogpolitik og sprogplanlægning i Europa. Artiklen giver et overblik over den information man kan finde om europæisk sprogpolitik og sprogplanlægning baseret på et omfattende spørgeskema med 9 sæt af spørgsmål. Data fra mere end 25 lande er tilgængelige via en webgrænseflade. Artiklen argumenterer for at adgangen til information om sprogpolitikker på tværs af landegrænserne er et vigtigt redskab for politikere, journalister, forskere og undervisere.

1. Introduction

The European Language Monitor (ELM) (Kirchmeier-Andersen et al. 2012) is a project of EFNIL, the European Federation of National Institutions for Language, that collects and publishes data on language legislation and language planning in Europe. ELM provides answers to questions like:

- How many countries have a language law?
- What languages are used as language of instruction in higher education?
- What languages are used on company web sites.
- Which countries have specific programs to support language technology for their languages?

The user can browse and compare language laws, find information about the status of minority languages, the use of languages in the educational systems, in the media and about the use of language technology in many European countries.

The data in ELM are collected and validated by the national institutions for language that are organized in EFNIL. The focus is on official regulations and their implementation. We have taken great care to provide comments, quotes, links and translations of legislation wherever possible. The data for ELM have been collected every 4 years since 2009. The current version, ELM 4, is based on data collected in 2017-2018.

2. What is a Language Monitor?

In some countries, such as Sweden and Norway, reports on the status of its language(s) are presented to policy makers on a regular basis, in others, language status reports are created ad hoc, depending on the political situation. Very few surveys are created on a regular basis allowing the comparison of language data across countries and over time.

In our view, a language monitor should comply with the following criteria:

- It is a scientific review of the language situation in one or more countries repeated in certain intervals.
- The information should be comparable over time.
- The information should be comparable across countries.

None of the three criteria are clear cut and easy to apply. It is not at all clear which kinds of data reflect the actual language situation of a country and which factors influence the change of that situation. Neither is it clear whether the data collected for one country are at all comparable to similar data from another, as the political and social conditions vary from country to country.

An important part of the data collection process, therefore, is to provide exact reference to the actual legislation in each country. A statistical overview and a maps view is provided as tools for further exploration and understanding of the different language regimes, their differences and similarities. We hope that this enables researchers and other interested parties to review the sources of the data, look into the details and draw their own conclusions.

The development of a language monitor is a continuous bootstrapping process where questions are tested and the answers evaluated, after which the questions are adjusted accordingly. In some cases, new questions are added. Thus in the newest version, ELM 4, a suite of questions on language technology has been added in cooperation with the META-NET project (Rehm & Uszkoreit 2012).

3. ELM vs. other Language Surveys

To our knowledge there are no surveys that cover the linguistic situation in Europe as detailed as ELM. The Unesco Survey of World Languages, which was launched in 2018, mainly focusses on the status and use of languages and especially endangered languages, but it does not provide quotes from and links to actual legislation, which is necessary if one wishes a more detailed view of the language policies in each country. Some of the EU Eurobarometers do contain questions about languages and linguistic practices, but they are based on public opinion data, not on legal facts. Other surveys, for instance The European Survey on Language Competences commissioned by CRELL (Araújo & Costa, 2013), focus on specific aspects and effects of language teaching and the role of external factors.

4. Methodology

ELM 4 contains 9 suites of questions. Some elicit lists of languages (What are the official languages in your country?). Some are simple yes/no questions (Is there a language law in your country?). Others offer multiple choices.

As many questions as possible have been designed to elicit quantifiable answers in order to give an overview. However, the most interesting and detailed information is located in the comment fields where links and quotes from legislation are provided (in the original language and in English) alongside with comments from EFNIL's language experts.

The question suites cover the following topics:

1. Country situation. Official, regional, indigenous, immigrant languages spoken within and outside the country, legal status, accordance with conventions
2. Legal situation. Language law, constitutional status, other regulations, language demands for citizenship
3. Primary and secondary education. Languages of instruction, languages offered
4. Tertiary education. Languages of instruction, languages used in publications and dissertations
5. Media. Papers, TV, film, music. Languages used and translations provided
6. Business. Regulations. Company languages, annual reports, websites
7. Dissemination of languages. Official languages taught abroad
8. Language organisations. Official, non--governmental but publicly funded, private
9. Language technology

The questionnaire was designed by EFNIL's ELM working group: Sabine Kirchmeier (Danish Language Council), Cecilia Robustelli, Academia della Crusca, Italy), Jennie Spetz (Swedish Language Council), Nina Teigland (Norwegian Language Council), Karlijn Watermanns (Nederlandse Taalunie), and presented and discussed with the EFNIL members. EFNIL representatives in all countries were asked to fill in the survey during 2017.

5. Visualisation and Translation

ELM is conceived as a transparent, interactive web-based system. This means that all questions and answers for all countries can be selected and displayed in a flexible manner. On the ELM website (ELM 4. 2019) it is possible to view the answers to all questions for a specific country, to compare the answers to a given question across countries and to combine questions and comments in order to get a more detailed picture. ELM also offers a map view of some of the data (currently only ELM 3. 2014).

Comments are given in English and quotes are given in the original language and in English translation. Active links to current legislation etc. are provided in most cases as shown in figure 1. Translations of the original quotes may be an authorized translation or provided by the respondent. This is indicated accordingly.

Fig. 1: ELM web interface showing responses to questions and comments in original language and in English.

6. Results

The following section presents some of the information that can be found using ELM. We will focus on two areas: Legal situation and language technology.

6.1 Legal situation

About half of the participation countries state that there are provisions about the official languages of the country stated in the constitution. Austria, Belgium, Bulgaria, Estonia, Finland, Hungary, Latvia, Lithuania, Portugal, Slovak Republic and Slovenia report that the official languages are stated in the constitution. The other countries do not have provisions of this kind.

Another way of securing a special status for the official languages of a country is a general legal act specifying the language use in various contexts. Only 6 countries do not have a specific language law.

Only Belgium, Greece, Luxembourg and Portugal do not report to have other legal acts regulating the use of official languages. However, all four have stated that they have specific language laws and for Belgium, Greece and Portugal, the use of official languages is also stated in the constitution.

In the following table, countries have been ranked according to the level of regulation they provide for their language(s). The least regulated countries, e.g. those with regulations at the lowest level of legislation, are on top, whereas those with the highest level of regulation appear at the bottom.

	Provisions for official languages		
Country	2.1 Constitution	2.2.Language law	2.3.Other legislation
Denmark	No	No	Yes
Germany	No	No	Yes
The Netherlands	No	No	Yes
Grand Duchy of Luxembourg	No	Yes	No
Greece	No	Yes	No
Czech Republic	No	Yes	Yes
Iceland	No	Yes	Yes
Norway	No	Yes	Yes
Sweden	No	Yes	Yes
UK	Not appl.	Not appl.	Yes
Portugal	Yes	No	No
Bulgaria	Yes	No	Yes
Hungary	Yes	No	Yes
Belgium	Yes	Yes	No
Austria	Yes	Yes	Yes
Estonia	Yes	Yes	Yes
Finland	Yes	Yes	Yes
Latvia	Yes	Yes	Yes
Lithuania	Yes	Yes	Yes
Slovak Republic	Yes	Yes	Yes
Slovenia	Yes	Yes	Yes

Table 1: Provisions for official languages at different levels of legislation

6.2 National Strategies and Funding Programs for Language Technology

Almost half of the countries report that they do have an official strategy for the development for their language(s). These may be protective strategies for minority languages, strategies for the support of languages in different situations, for the choice of language of instruction at universities and schools, or strategies for how languages can be supported technologically.

The existence of targeted funding programs can be seen as an indication for the awareness in a country about the importance of language technology. More than half of the countries have funding programs for LT. These are mainly the Nordic and East European countries, Luxembourg and the UK, but also Greece, Germany and Hungary.

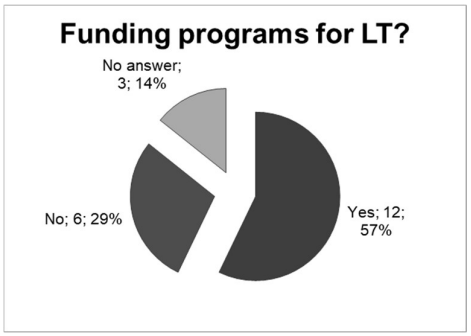


Fig. 2. More than half of the countries have dedicated funding programs for Language Technology.

7. More Details on Current Language Legislation with ELM and LLE

As described above, ELM provides detailed statistical information about the status of the official and non-official languages in Europe and the different rulings and practices in European countries. Supplementary to the statistical information, the comment fields for each question contain a wealth of details about each specific country and inspiration for further studies, such as links to information sources, quotes from legislation texts and their translations into English.

In addition to the ELM project, EFNIL also provides a detailed description of the language legislation in effect in each country in the form of an overview article for each country in both French and English. The project, called LLE (Language Legislation in Europe), is also based on the information provided by the European national institutions for language and frequently updated.

8. Conclusion

Statistical information about language legislation and practices is highly relevant for decision makers, journalists, researchers and teachers. Comparing the different approaches to language legislation and planning in Europe generates new perspectives on local practices. However, the statistical data alone cannot capture the diversity that is reflected in the different language regimes, and sometimes the statistics may raise more questions than they answer.

EFNIL's two projects ELM (European Language Monitor) and LLE (Language Legislation in Europe) provide a unique, reliable and easily accessible source of information on the language situation in the participating countries and allow the user to explore in detail the complexity of the language situation in Europe.

9. Bibliographical References

Araújo, L. and and Dins da Costa, P. (2013): *The European Survey on Language Competences.* Commissioned by CRELL Centre for Research on Education and Lifelong Learning. JRC 82366.

Kirchmeier-Andersen et al. (2012): Sabine Kirchmeier-Andersen, Cecilia Robustelli, Jennie Spetz, Gerhard Stickel, Nina Teigland. European Language Monitor (ELM). In Stickel, Gerhard: *National, Regional and Minority Languages in Europe. Contributions to the Annual Conference 2009 of EFNIL in Dublin.* Duisburger Arbeiten zur Sprach- und Kulturwissenschaft / Duisburg Papers on Research in Language and Culture. Berlin. Peter Lang, pp 181-190.

Rehm, G. and Uszkoreit, H., editors. (2012). META-NET White Paper Series: Europe's Languages in the Digital Age. Springer, Heidelberg, New York, Dordrecht, London. 31 volumes on 30 European languages. http://www. meta-net.eu/whitepapers.

ELM 4 (2019). European Language Monitor. Experimental version: https://juniper.nytud.hu/elm4

ELM 3 (2014). European Language Monitor. Articles and general information: http://efnil.org/projects/elm

Language Legislation in Europe (LLE): http://efnil.org/projects/lle

Proceedings, pages 335–337
Paris, UNESCO Headquarters, 5-6 December, 2019.

Achieving the Goal of Language Technology for All

Ramakrishnan A G
MILE Laboratory, Department of Electrical Engineering
Indian Institute of Science, Bangalore, India
agr@iisc,ac.in

Abstract

Given the advances in information technology, communication and internet, enabling individuals, organizations and governments to carry out their day-to-day transactions without being constrained in any way by the multitude of living languages is completely feasible, provided there is commitment to the cause and constant, undaunted efforts. In a country like India, this must be an ongoing process, since the various fields of knowledge are constantly advancing, giving rise to new terminologies and challenges. It requires systematic planning and execution by standing committees, both at the national and state levels, who work together, constantly communicating and collaborating with one another. Further, it may also be worthwhile looking at establishing a national level research organization for continuous upgrading of speech and translation technologies, especially in a code-mixed scenario.

Keywords: standardization, standing committees, translation, commitment, vision, policies, multilinguality, fonts, web content, multilingual dictionaries, open databases, transcription, computational linguistics.

Résumé

பல வாழும் மொழிகள் இருந்தாலும், தகவல் தொழில்நுட்பம், தகவல் தொடர்பு மற்றும் இணையம் ஆகியவற்றின் சிறந்த முன்னேற்றங்களால், தனிநபர்கள், நிறுவனங்கள் மற்றும் அரசாங்கங்கள் தங்கள் அன்றாட பரிவர்த்தனைகளை எந்தவொரு வகையிலும் கட்டுப்படாமல், குறையின்றி செயல்படுத்துவது முற்றிலும் சாத்தியமானதே. ஆனால், அதற்கான அர்ப்பணிப்பும், நிலையான, இடைவிடாத முயற்சிகளும் தேவை. இந்தியா போன்ற ஒரு நாட்டில், இது ஒரு இடைவிடாது எப்போதும் தொடரும் செயல்முறையாக இருக்க வேண்டும், ஏனென்றால் பல்வேறு அறிவுத் துறைகள் தொடர்ந்து முன்னேறி வருகின்றன. இது புதிய சொற்களஞ்சியங்களுக்கும், சவால்களுக்கும் வழிவகுக்கிறது. இதற்கு, தேசிய மற்றும் மாநில மட்டங்களில், ஒருவருக்கொருவர் தொடர்ந்து தொடர்புகொண்டு, ஒன்றிணைந்து செயல்பட்டு, ஒத்துழைக்கும் நிலைக்குழுக்களும், அவற்றின் முறையான திட்டமிடல் மற்றும் செயல்படுத்தலும் தேவைப்படுகிறது.

1. Preamble

Even though more than 70 years have passed by since independence, none of the living languages, including the 22 scheduled languages, have kept pace with the scientific and technological advances and the consequent new terminologies. Thus, it is not possible for anyone, even if he/she is really interested, to pursue higher studies in science and technology in any of the Indian languages, including Hindi, which is being spoken by the majority.

The author feels that there has been no systematic planning to develop any of these languages to keep pace with the time. However, it is still not too late, since only in the recent past, the information and communication technologies have been advancing fast. Thus, with proper planning, commitment and systematic execution, India can still catch up and develop language technologies in all the major living languages and enable individuals, organizations and governments to improve their performance levels in spite of the presence of multiple, interacting and intersecting languages. It must also be realized that this must be planned as an eternal activity, and not as something that will get completed within a particular time.

1.1 Need for Multipronged Effort

There are multiple steps that need to be taken simultaneously and pursued vigorously. This involves standardization of terminologies in each language, technologies and the interfaces between themselves and other application softwares in multiple domains, wherein, again new developments will keep taking place.

1.2 Need for Multiple Standing Committees

Each state in India must have a standing committee, which consists of domain experts from different fields, linguists and technologists, who will periodically meet, plan and advise the respective arms of the industry and Government as to the steps taken for the next level of improvement of language technologies, newer applications, their deployment and outreach.

1.3 Ongoing Standardization of Terminology

To begin with, a systematic procedure must be evolved, which specifies the approach to be followed in coining a word for a new idea or object or an action. If possible, for languages such as Bangla and Tamil, which are official languages in countries other than India also, there can be understanding between the countries with respect to this activity, or even the committee may be constituted with members from all those countries. The committee for standardization of special terms to be used in business, art, science, technology, management and other special fields must periodically be creating new terms in the respective languages, publish them in dedicated websites for feedback from the experts and user community (people in general) for a specified length of time and then announce them as standard.

The members of the above committee may change with time; however, the committee itself needs to be planned as an ongoing, never-ending structure.

1.4 Promoting the Regular Use of Indian Languages by School Students

Each state must legislate that every QWERTY keyboard must have the local state script also printed or painted in the keys. Further, both the state educational departments and other interested organisations can conduct yearly competitions in the schools for fast typing in Indian languages using keyboards optimized for that language. By making the prizes attractive, we can easily ensure that the next generation is comfortable in typing in Indian languages using efficient key inputs, rather than using phonetic keyboards with input in Roman script.

1.5 Creating more Indian Language Web Content

Once again, attractive incentives may be given to school, college students as well as general public for creating quality web content in Indian languages. One simple way is to translate English Wikipedia content, as well as the content in Government tourism websites, etc. into the local language.

1.6 Role of Institution of Engineers, IETE, etc.

Professional bodies such as the Indian Academy of Sciences, Indian National Science Academy, Institution of Engineers (India), Institution of Electronics and Telecommunication Engineers can play a very significant role in promoting all the above. More importantly, they can announce awards for the best projects in undergraduate, postgraduate and even doctoral level work that create new applications in Indian languages, or promote the creation and widespread use of Indic language content. They must also play an important role in the policy making in these important areas.

2. Standardization of Fonts for Display Boards

With the existence of multiple languages, and with expanding travel of people for both business and tourism, there must be development of focused technology development, as well as standardization of different kinds, so that any traveler is facilitated to easily navigate through any place in spite of the display boards being in an unknown language or script. With camera captured document image analysis and recognition technologies, it is now possible to automatically detect and extract the text from images captured by a mobile phone. The suggestion is to standardize one or two fonts and even font sizes for any public display boards, so that the recognition engines can be optimized for these fonts and font sizes. By also standardizing the colour of the text and the background wherever possible, we can significantly increase the text recognition performance, making it a technology usable on a daily basis. The recognized text can then be translated and easily understood by the traveler. In many circumstances, most of the words (at least the key ones) in such boards will be proper nouns and hence, even a transliteration or transcription in the target script may suffice.

The same suggestion is also advanced for all the official printed documents of the government, which will guarantee exceptional accuracy of the respective OCRs, simplifying the process of digitization, editing or updating of existing documents, wherein the source e-text is not easily available or accessible.

3. Create Specialized Institutes for Language Technology

New institutions must be created, whose mandate is to primarily research and develop new language technologies, and also train people for the industry.

3.1 Both Hardware and Software

These institutions will look at both hardware and software aspects. Hardware includes design of new devices, both primary and peripherals, such as input and output. For example, an extremely desirable device for a multilingual country such as India is a handheld, dedicated, handwriting input device, which will wirelessly transmit the recognized text to any computing device available nearby. The same universal device can be used for different languages by different people, by downloading the appropriate recognition engine, may be after a fee paid online. This has the potential to become as ubiquitous as the QWERTY keyboard, if not more.

3.2 Creating Computational Linguistic Studies

The computer science curriculum in all the engineering institutions must necessarily involve study of natural language processing and basics of computational linguistics. Also, the curricula for the different specialties in linguistics in universities must ensure basic training of the graduates in computer applications and computational linguistic tools. Research leading to up to Ph D degree in computational linguistics must be introduced in higher educational institutions.

4. Designing for Multilinguality

4.1 Website Design

All websites must necessarily be designed to be multilingual. The design specifications must take into account issues such as standardized Tables in each of the languages, wherever lists are involved, and provision for the users to quickly and periodically update such lists, without requiring technical experts to carry out such tasks. Once multilinguality becomes a basic, mandatory feature of websites, such an updation will be regularly required.

4.2 Design for Ease of Adding Another Language for User Interface

While the internal representations may be in English, the graphical user interface must be in the local language. In fact, it is preferable to make a provision for the user to customize the user interface to the language of his choice. This must be designed in a way that a technical expert is not required to add the new language interface.

4.3 Standardization of Terms for Lists

The terms for the most common lists that can occur in popular applications must be standardized for all the scheduled languages and a Table, giving the lists of corresponding terms across all the languages must be published and made openly available. This is because, in the past, there have been many efforts, both by individual

groups and many industries to come out with such terms in Indian languages, which led to the existence of multiple, unrelated words for the same object, adding to the confusion.

4.4 Creating Multilingual Digital Dictionaries

Structured digital lists containing equivalent verb roots, verb phrases, adjectives, adverbs, common nouns and noun phrases must be created, standardized and openly available for software developers. Laws must be created, which make it mandatory to use these standardized lists in all applications. These multilingual dictionaries will also form an extremely useful component of the machine translation. Further, they may ensure that information is not lost, when a sentence or even a document is translated between a number of successive pairs of languages.

5. Policy Changes Needed

5.1 Making Research Data Easily Available

Every time an academic institute is funded to develop any aspect related to language technology, there must be a clear condition that the data collected and may be annotated as part of the funded project must be made available to researchers in some standard format, as soon as the project term ends, unless the researcher starts an industry or transfers the technology to an industry. In the latter case, there can be appropriate new policies that are applicable.

There could be a national level standing committee, which can look into these policies and the changes required from time to time.

5.2 Publications Arising out of Government Funding being Freely Available for Researchers

Just as special copyright laws exist in USA for Government funded research, exclusive copyright laws must be enacted, by which the publications arising out of any Government funded research must be openly available and/or the copyright must rest with the Government or the researcher. In any case, the idea is to ensure that the results of public funded research are readily available to the research community, without having to again pay to access them.

5.3 Synergy between Multiple Institutions

Currently, there are many planning, policy making and funding agencies, whose mandates or activities overlap; however, many of them operate in silos. This may lead to inefficient use of public money, and outcomes that fall short of desirable performance or quality standards or expectations. The national level standing committee must also look into such matters and advise the government as to how to bring in synergy between such agencies.

6. Conclusion

Several suggestions have been made to the researchers, funding agencies and the Governments, which, in the opinion of the author, will go a long way in reaching the benefits of language technology to one and all.

There are a number of other things that need to be considered and this article is by no means exhaustive. However, if the spirit of the article is understood, one can come up with meaningful suggestions to tackle each and every issue not addressed here.

7. Acknowledgements

The author gratefully acknowledges Tata Trust Travel Grant for funding him to travel and participate in this conference. Immense thanks are also due to the Technology Development for Indian Languages (TDIL), Ministry of Information Technology, Government of India, for funding many of his projects in language technology, which has made it possible for him to be invited to this conference. Acknowledgment is also due to many students, research staff, colleagues, collaborators and interns, who enriched his knowledge and experience.

Proceedings, pages 338–341
Paris, UNESCO Headquarters, 5-6 December, 2019.

Semi-supervised Learning by Machine Speech Chain for Multilingual Speech Processing, and Recent Progress on Automatic Speech Interpretation

Satoshi Nakamura, Sakriani Sakti, Katsuhito Sudoh
Graduate School of Science and Technology, Nara Institute of Science and Technology
8916-5, Takayama Ikoma, Nara, 630-0192,Japan
{s-nakamura, ssakti, sudoh}@is.naist.jp

Abstract

In this paper, we introduce our recent machine speech chain frameworks based on deep learning that learned, not only to listen or speak but also listen while speaking. This is the first deep learning model that integrates human speech perception and production behaviors. First, we describe the primary machine speech chain architecture that integrates automatic speech recognition (ASR) and text-to-speech synthesis (TTS). After that, we describe the use of machine speech for code-switching ASR and TTS. Also, this paper describes our attempts to automatic simultaneous machine interpretation. Finally, we discuss the possibility and difficulty.

Keywords: Speech Recognition, Speech Synthesis, Speech Chain, Machine Translation, Speech Interpretation, Deep Learning

1. Introduction

Many attempts have been made to replicate human speech perception and production by machines. To date, the development of advanced spoken language technologies based on ASR and TTS has enabled computers to either learn how to listen or speak. However, despite the close relationship between speech perception and production, ASR and TTS researches have progressed more or less independently without exerting much mutual influence on each other. Thus constructing ASR or TTS is commonly done in supervised fashion; a large amount of paired speech and corresponding transcription are used. However, paired data is not always available for under-resourced languages, code-switching situation, and cross-language situations. In this paper, we first introduce semi-supervised learning by Machine Speech Chain. Then we describe our attempts to automatic simultaneous machine interpretation.

2. Machine Speech Chain

By simultaneously listening and speaking, the speaker can monitor her volume, articulation, and the general comprehensibility of her speech. Therefore, a closed-loop speech chain mechanism with auditory feedback from the speaker's mouth to her ear is crucial.

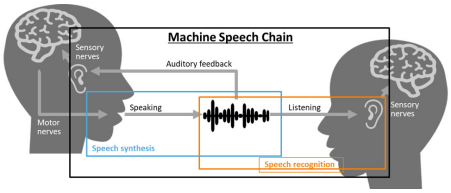


Figure 1: Speech chain.

The speech chain, which was first introduced by Denes et al. (Denes and Pinson, 1993), described the basic mechanism involved in speech communication when a spoken message travels from the speaker's mind to the listener's mind (Fig. 1). It consists of a speech production mechanism in which the speaker produces words and generates speech sound waves, transmits the speech waveform through a medium (i.e., air), and creates a speech perception process in a listener's auditory system to perceive what was said.

We have introduced machine speech chain frameworks based on deep learning that learned, not only to listen or speak but also listen while speaking. To the best of our knowledge, this is the first deep learning model that integrates human speech perception and production behaviors. The framework allows us to perform semi-supervised learning and avoids the need for a large amount of paired speech and text data. Specifically, the structure enables ASR and TTS to assist each other when they receive unpaired data since it allows them to infer the missing pair and optimize the models with reconstruction loss. First, we describe the primary machine speech chain architecture that integrates ASR and TTS. After that, we describe the use of machine speech for code-switching ASR and TTS.

An overview of our proposed machine speech chain architecture is illustrated in Fig. 2. It consists of a sequence-to-sequence ASR (Bahdanau et al., 2016), a sequence-to-sequence TTS (Wang et al., 2017), and a loop connection from ASR to TTS and from TTS to ASR. The key idea is to jointly train both the ASR and TTS models. As mentioned above, the sequence-to-sequence model in closed-loop architecture allows us to train our model on the concatenation of both the labeled and unlabeled data. For supervised training with labeled data (the speech utterances $\mathbf{x}$ and the corresponding text transcription $\mathbf{y}$ from dataset $\mathcal{D}^P$), both ASR and TTS models can be trained independently by minimizing the loss between their predicted target sequence and the ground truth sequence (calculating $\mathcal{L}_P^{ASR}$ for ASR and $\mathcal{L}_P^{TTS}$ for TTS).

However, for unsupervised training with unlabeled data

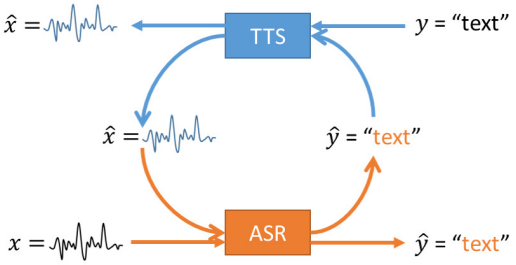


Figure 2: Overview of machine speech chain architecture by deep learning.

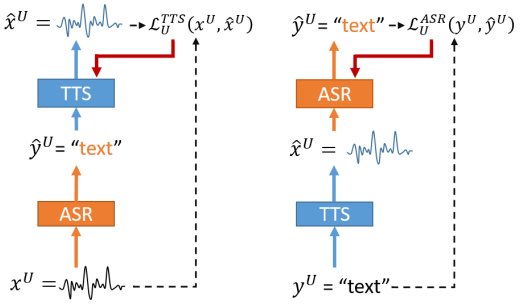


Figure 3: Examples of unrolled process in machine speech chain: (a) from ASR to TTS and (b) from TTS to ASR.

(speech only or text only), both models need to support each other through a connection. To further clarify the learning process during unsupervised training, we unrolled the architecture as follows:

- **Unrolled process from ASR to TTS**
  Given only the speech utterances $\mathbf{x}$ from unpaired dataset $\mathcal{D}^U$, ASR generates the text transcription $\hat{\mathbf{y}}$. TTS then reconstructs the speech waveform $\hat{\mathbf{x}}$ given the generated text $\hat{\mathbf{y}}$ from ASR and calculate the loss $\mathcal{L}_U^{TTS}$ between $\mathbf{x}$ and $\hat{\mathbf{x}}$. Fig. 3(a) illustrates the mechanism. We may also treat it as an autoencoder model, where the speech-to-text ASR serves as an encoder and the text-to-speech TTS as a decoder.

- **Unrolled process from TTS to ASR**
  Given only the text transcription $\mathbf{y}$ from unpaired dataset $\mathcal{D}^U$, TTS generates speech waveform $\hat{\mathbf{x}}$, while ASR also reconstructs the original text transcription $\hat{\mathbf{y}}$ given the synthesized speech $\hat{\mathbf{x}}$. After that, we calculate the loss $\mathcal{L}_U^{ASR}$ between $\mathbf{y}$ and $\hat{\mathbf{y}}$. Fig. 3(b) illustrates the mechanism. Here, we may also treat it as another autoencoder model, where the text-to-speech TTS serves as an encoder and the speech-to-text ASR as a decoder.

We combine all loss together and update both ASR and TTS model:

$$\mathcal{L} = \alpha * (\mathcal{L}_P^{ASR} + \mathcal{L}_P^{TTS}) + \beta * (\mathcal{L}_U^{ASR} + \mathcal{L}_U^{TTS}) \quad (1)$$

where α, β are hyper-parameters to scale the loss between supervised (paired) and unsupervised (unpaired) loss.

Experiment

Table 1: Experiment result for single-speaker test set.

Data	Hyperparameters			ASR	TTS	
	α	β	gen. mode	CER (%)	Mel	Raw
Paired (10k)	-	-	-	10.06	7.068	9.376
+ Unpaired (40k)	0.25	1	greedy	5.83	6.212	8.485
	0.5	1	greedy	5.75	6.247	8.418
	0.25	1	beam 5	5.44	6.243	8.441
	0.5	1	beam 5	5.77	6.201	8.435

We utilized both monolingual Japanese and English ATR Basic Travel Expression Corpus (BTEC) (Kikui et al., 2003) which cover the basic conversations in the travel domain, such as sightseeing, restaurant, hotel stays, etc.

Single-speaker Monolingual Task

To gather a large single speaker speech dataset, we utilized Google TTS[1] to generate a large set of speech waveform based on monolingual English BTEC sentences.

Table 1 shows our result on the single-speaker ASR and TTS experiments. For the ASR experiment, we used a character error rate (CER) for evaluating the ASR model. For the TTS experiment, we reported the MSE between the predicted log Mel and the log magnitude spectrogram to the ground truth. We also report the accuracy of our model that predicted the last speech frame. We used different values for α and text decoding strategy for ASR (in the unsupervised learning stage) with a greedy search or a beam search.

The results show that after ASR and TTS models have been trained with a small paired dataset, they start to teach each other using unpaired data and generate useful feedback. Here we improved both ASR and TTS performance. Our ASR model reduced CER by 4.6% compared to the system that was only trained with labeled data. In addition to ASR, our TTS also decreased the MSE and the end of speech prediction accuracy.

3. Speech chain for code-switching speech

Here we discuss the possibility to apply the machine speech chain framework for code-switching (CS) ASR and TTS tasks.

CS speech, in which speakers alternate between two or more languages in the same utterance often occur in multilingual communities. The common way of developing spoken language technologies for code-switching relies on a supervised manner that requires a significant amount of CS data to train the models. unfortunately, parallel speech and transcription of CS data suitable for training ASR and TTS are mostly unavailable.

We propose to utilize the machine speech chain framework to enable code-switching ASR and TTS training in semi-supervised fashion. In particular, we construct with following learning process:

- **Train ASR and TTS separately with parallel speech-text monolingual data (supervised learn-**

[1]Google TTS – https://pypi.python.org/pypi/gTTS

ing) We first separately train the ASR and TTS systems with parallel speech-text of monolingual data (supervised learning). Given a speech and text pair of monolingual data (x^{Mono}, y^{Mono}) with speech length S and text length T, ASR generates text probability vector $\hat{y}^{Mono}$ using teacher-forcing, and loss $\mathcal{L}^{ASR}_{Mono}(\hat{y}^{Mono}, y^{Mono})$ is calculated between output text probability vector $\hat{y}^{Mono}$ and reference text y^{Mono}. On the other hand, TTS also generates best predicted speech $\hat{x}^{Mono}$ using teacher-forcing, and loss $\mathcal{L}^{TTS}_{Mono}(\hat{x}^{Mono}, x^{Mono})$ is calculated between predicted speech $\hat{x}^{Mono}$ and ground-truth speech x^{Mono}. The parameters are then updated with gradient descent optimization.

- **Train ASR-TTS simultaneously in a speech chain with unparallel CS data (unsupervised learning)** After that, we then simultaneously train ASR and TTS through a speech chain with unparallel CS data (unsupervised learning).

To further clarify the learning process during unsupervised training, we unrolled the following architecture:

- **Unrolled process from TTS to ASR given only CS text**
  Given CS text input y^{CS} only, TTS generates speech waveform $\hat{x}^{CS}$, while ASR also attempts to reconstruct original text transcription $\hat{y}^{CS}$, given the synthesized speech. Then loss $\mathcal{L}^{ASR}_{CS}(\hat{y}^{CS}, y^{CS})$ can be calculated between output text probability vector $\hat{y}^{CS}$ and input text y^{CS} to update the ASR parameters.

- **Unrolled process from ASR to TTS given only CS speech**
  Given unlabeled CS speech features x^{CS}, ASR transcribes unlabeled input speech $\hat{y}^{CS}$, while TTS attempts to reconstruct original speech waveform $\hat{x}^{CS}$ based on the output text from ASR. Then loss $\mathcal{L}^{TTS}_{CS}(\hat{x}^{CS}, x^{CS})$ can be calculated between reconstructed speech waveform $\hat{x}^{CS}$ and the input of original speech waveform x^{CS} to update the TTS parameters.

Experiment

We randomly selected 50k sentences for training, 500 sentences for the development set, and 500 sentences for test set from BTEC1-4. As large Japanese-English CS data do not exist yet, we constructed it from monolingual Japanese and English BTEC sentences. Here, we created two types of intra-sentential code-switching: word-level and phrase-level code-switching and phrase-level. For more detail can be found in (Nakayama et al., 2018a). Here we also utilized Google TTS to generate speech from the text corpora.

Here, we use "Ja25k+En25k" baseline system which is ASR or TTS that was trained in supervised learning with 25k monolingual Japanese text and the corresponding speech plus 25k monolingual English text and the corresponding speech.

Table 2 shows ASR-TTS performances (in CER and L2-norm squared, respectively) of the baseline and the proposed CS speech chain framework that was trained in semi-supervised fashion using monolingual Ja25k+En25k as paired data and code-switching CSWord+Phr as unpaired data.

Our proposed speech-chain model could significantly improve the ASR system in CS test set TstCSWord+ Phr from 18.11% CER down to 5.35%, while keeping the good performance in monolingual setting (only slightly CER reduction up to 0.1% and 0.7% for Japanese and English monolingual test set, respectively). The same tendency is also shown in TTS results. It could also improve the TTS system in CS test set TstCSWord+Phr from 0.489 to 0.374 L2-norm squared while keeping similar performance for Japanese and English monolingual test set. For more detail can be found in (Nakayama et al., 2018b).

Table 2: ASR & TTS performances (in CER & L2-norm squared, respectively) of the proposed CS speech chain framework.

TstMonoJa		**TstCSWord+Phr**		**TstMonoEn**	
ASR	**TTS**	**ASR**	**TTS**	**ASR**	**TTS**
Baseline: Supervised training Ja25k+En25k (Monolingual, speech-text paired data)					
1.71%	0.312	18.11%	0.489	2.99%	0.437
Speech chain: Semi-supervised training +CSWord+Phr10k (CS, unpaired data)					
1.81%	0.312	5.35%	0.374	3.69%	0.437

4. Speech-to-speech Translation

Speech-to-speech translation (S2ST) technology is key for cross-lingual communication. However, there have been various technical difficulties and difficulties in collecting paired data of source and target language speech and text corpora. S2ST in Japan had been started to overcome the language barrier problem in 1986. So far, we have been working on speech recognition, machine translation, speech synthesis and integration for an S2ST system. S2ST between Western languages and a non-Western language, such as English-from/to-Japanese, or English-from/to-Chinese, requires technologies to overcome the drastic differences in linguistic expressions. For example, a translation from Japanese to English requires, (1) a word separation process for Japanese because Japanese has no explicit spacing information, and (2) transforming the source sentence into a target sentence with a drastically different style because their word order and their coverage of words are completely different, among other factors. The overall speech-to-speech translation system is shown in Fig. 4. The system consists of three major modules, i.e., a multilingual speech recognition module, a multilingual machine translation module, and a multilingual speech synthesis module.

In addition to the End-to-end ASR and TTS, End-to-end machine translation algorithms such as the encoder-decoder with attention (Luong et al., 2015) realized high-performance MT these days. However, the current approach is far from human interpreters in (1) simultaneity, (2)transfer of para-linguistic information of emotion and

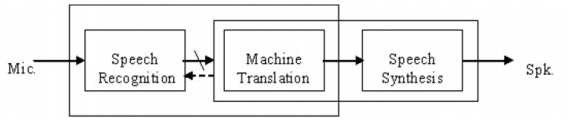


Figure 4: Block diagram of the S2ST system

emphasis, (3)dialog and multi-modal situation contexts and mutual grounding, (4) differences of cultural backgrounds, (5)interpreting intentions.

For simultaneity, conventional speech translation systems wait until the end of the input sentence before starting translation, causing a large delay in the translation process. Previous methods have been proposed to reduce this delay by dividing the input utterance on pause boundaries, but while these methods have proven useful on speech translation of language pairs with similar word order, they are insensitive to linguistic information and less effective for languages that require more word reordering. We have been proposed two approaches. The first one is to use the phrase table and reordering probabilities used in phrase-based translation systems to decide points in the sentence where we can begin translation with less delay (Fujita et al., 2013). The second one is to apply syntax-based SMT to simultaneous translation, and propose two methods to prevent accuracy degradation: a method to predict unseen syntactic constituents that help generate complete parse trees and a method that waits for more input when the current utterance is not enough to generate a fluent translation (Oda et al., 2015)

For the transfer of para-linguistic information of emphasis, we have been proposed a method based on encoder-decoder with attention (Do et al., 2018). This method estimates emphasis in the source speech and map into target speech within encoder-decoder cascaded speech-to-speech translation framework. This framework will be extended to incorporate emotions in future. Another attempt is to realize direct speech-to-speech translation to translate linguistic and para-linguistic information into one framework. We have been proposed a method using colloquium training based on encoder-decoder direct speech translation (Kano et al., 2017).

5. Conclusion

This paper demonstrated recent developments in machine speech chain mechanism based on deep learning and automatic speech interpretation. The machine speech chain mechanism will be useful for building ASR and TTS for under-resourced languages. We only described single speaker results but for the multi-speaker situation, details can be found in (Tjandra et al., 2017; Tjandra et al., 2018; Tjandra et al., 2019). For S2ST there still remain many research problems for real cross-lingual natural communication. However, these speech and language technologies will contribute not only cross-lingual communication but preservation of spoken languages.

6. ACKNOWLEDGEMENTS

Part of this work was supported by JSPS KAKENHI Grant Numbers JP17H06101 and JP17K00237.

Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P., and Bengio, Y. (2016). End-to-end attention-based large vocabulary speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 4945–4949. IEEE.

Denes, P. and Pinson, E. (1993). *The Speech Chain*. Anchor books. Worth Publishers.

Do, Q. T., Sakti, S., and Nakamura, S. (2018). Sequence-to-sequence models for emphasis speech translation. *IEEE/ACM Trans. Audio, Speech & Language Processing 26(10)*, pages 1873–1883.

Fujita, T., Neubig, G., Sakti, S., Toda, T., and Nakamura, S. (2013). Simple, lexicalized choice of translation timing for simultaneous speech translation. In *Proc. of INTERSPEECH*, pages 3487–3490, Lyon, France.

Kano, T., Sakti, S., and Nakamura, S. (2017). Structured-based curriculum learning for end-to-end english-japanese speech translation. In *Proc. of INTERSPEECH*, pages 2630–2634, Stochholm, Sweden.

Kikui, G., Sumita, E., Takezawa, T., and Yamamoto, S. (2003). Creating corpora for speech-to-speech translation. In *Proc. of EUROSPEECH*, pages 381–384, Geneva, Switzerland.

Luong, M.-T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *CoRR*.

Nakayama, S., Kano, T., Do, Q. T., Sakti, S., and Nakamura, S. (2018a). Japanese-english code-switching speech data construction. In *Proc. of Oriental COCOSDA*, pages 67–71, Miyazaki, Japan.

Nakayama, S., Tjandra, A., Sakti, S., and Nakamura, S. (2018b). Speech chain for semi-supervised learning of japanese-english code-switching asr and tts. In *Proc. of IEEE SLT*, pages 182–189, Athen, Greece.

Oda, Y., Neubig, G., Sakti, S., Toda, T., and Nakamura, S. (2015). Syntax-based simultaneous translation through prediction of unseen syntactic constituents. In *Association for Computational Linguistics*, pages 198–207, Beijing, China.

Tjandra, A., Sakti, S., and Nakamura, S. (2017). Listening while speaking: Speech chain by deep learning. In *Proc. of IEEE ASRU*, pages 301–308, Okinawa, Japan.

Tjandra, A., Sakti, S., and Nakamura, S. (2018). Machine speech chain with one-shot speaker adaptation. In *Proc. of INTERSPEECH*, pages 887–891, Hyderabad India.

Tjandra, A., Sakti, S., and Nakamura, S. (2019). End-to-end feedback loss in speech chain framework via straight-through estimator. In *Proc. of ICASSP*, pages 6281–6285, Brighton, UK.

Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R., and Saurous, R. A. (2017). Tacotron: A fully end-to-end text-to-speech synthesis model. In *Proc. of INTERSPEECH*, pages 4006–4010, Stockholm, Sweden.

Proceedings, pages 342–345
Paris, UNESCO Headquarters, 5-6 December, 2019.

Parsing the Less-configurational Georgian Language with a Context-Free Grammar

Oleg Kapanadze
Tbilisi State University
1, Chavchavadze Ave.,
0179 Tbilisi
Georgia
okapanadze@uni-potsdam.de

Abstract

A large part of the methodology for Natural Language Processing has been developed for languages with a strong syntactic configuration. At the other end of the configurational spectrum there are languages with rich derivational and inflectional morphology. These languages for *morphologically rich and less-configurational* features are referred to as **MR&LC**. In our study we have addressed Georgian - a language with less-configurational constraints, though, with a rich inflectional morphology and a very little fixed structure on the sentence level, and therefore, the most syntax-level information for the Georgian language is conveyed by its productive morphology.
This paper features issues concerned with development of a crucial NLP resource for the Georgian language - a Context-Free Syntactic Parser.

Keywords: Georgian Language Processing, Morphologically rich and less-configurational languages, Context-Free Syntactic Parser

Résumé

ბუნებრივი ენების ტექნოლოგიისათვის განკუთვნილი მეთოდოლოგიის უდიდესი ნაწილი შემუშავებულია სინტაქსური კონფიგურაციაზე მკაცრი შეზრუდვების მქონე ენებისთვის. კონფიგურაციული სპექტრის საპირისპირო მხარეს წარმოდგენილია ენები, რომლებიც ნაყოფიერი დერივაციული და ფლექსიური მორფოლოგიით გამოირჩევიან. ასეთი ტიპის ენებს მათი ფართო მორფოლოგიური შესაძლებლობებისა და სინტაქსური თვალსაზრისით ნაკლებად შეზღუდულობის გამო *მდიდარი მორფოლოგიისა და ნაკლებად კონფიგურირებულ* - მმ&ნკ ენებს უწოდებენ. ამ კუთხით ჩვენი კვლევის საგანია ქართული - ენა, რომელის სტრუქტურა მცირე კონფიგურაციული შეზღუდვებით გამოირჩევა და მდიდარი ფლექსიური მორფოლოგიის წყალობით წინადადების დონეზე ნაკლებად ფიქსირებული სინტაქსური სტრუქტურებით არის წამოდგენილი. ამავდროულად ქართული ენის წინადადების სინტაქსური სტრუქტურის შესახებ ინფორმაცია ზმნის ვალენტობისა და მისი შესაბამისი მორფოლოგიური მარკერების საშუალებით არის ხელმისაწვდომი.
წინამდებარე სტატიაში განიხილულია საკითხები, რომლებიც ეხება ქართული ენის ტექნოლოგიისათვის უმნიშვნელოვანესი რესურსის - კონტექსტისაგან დამოუკიდებელი სინტაქსური პარსერის შემუშავებას.

1. Introduction

A large part of the methodology for natural language processing (NLP) has been developed for English which is known as a strongly configurational language. Hence, nearly all the syntactic information needed by any NLP application for English can be obtained by configurational analysis. At the other end of the configurational spectrum are the languages with rich derivational and inflectional morphology, such as Georgian that has very little fixed structure on the sentence level. These languages for *morphologically rich and less-configurational* features are referred to as **MR&LC (**Fraser et al., 2001). All of them are thriving to get a place in the modern digital world and in order to profit of the new opportunities offered by the Internet and digital devices must be modeled for using in high-quality computing systems. The long-term viability of languages not specifically supported by Human Language Technology is therefore put at risk and they can seriously face digital extinction.

There are a multitude of academic grammars and dictionaries developed for the Georgian language. owever, this does not mean that there is a sufficient support for computational applications involving Georgian, as these resources are not suited for NLP needs.

The proposed presentation will feature issues concerned with the development of a crucial NLP resource — a syntactic parser for the Georgian language. To this end we used a methdology that will extract a FST grammar and a consequent lexicon from a monolingual Georgian TreeBank. The compiled language resources will be utilized for the Georgian text syntactic annotation which terminal nodes are saturated with rich morphologic features.

2. Treebanking in NLP

A monolingual ***TreeBank*** is a parsed corpus in which sentences are annotated with syntactic structure. They are skeletal parses showing syntactic information – a ***bank*** of linguistic ***trees***. Syntactic structure is commonly represented as ***a tree structure*** (in Mathematical terms – ***an Oriented Graph*****)**, hence the name ***TreeBank*****.**

TreeBanks have become valuable resources as repositories for linguistic research, since **c**orpus-based methods became useful in multilingual lexicography playing an important role in empirical language studies. They can be used in *languages contrastive studies* and *translation science*, in *corpus linguistics* for studying syntactic phenomena, in computational linguistics as evaluation corpora for different Human Language Technology systems or for training and testing *parsers* and as a database for *Translation Memory* systems.

TreeBanks can be created completely manually or semi-automatically, where a parser assigns some syntactic structure to a text that is then checked by linguists and, if necessary, corrected. Treebanks are often created on top of a corpus that has already been annotated with part-of-speech tags. The annotation can vary from constituent to dependency or tecto-grammatical structures. Additionally, treebanks are sometimes enhanced with semantic or other linguistic information.

Some TreeBanks follow a specific linguistic theory (e.g. the Bulgarian language follows HPSG), but most try to be less theory-specific. However, two main groups can be distinguished: treebanks that annotate *phrase structure* (the *Penn TreeBank* for Arabic, English and Chinese) and those that annotate *dependency structure* (the *Prague Dependency TreeBank* for the Czech language).

A significant part of modern treebanking literature is devoted to creation of large TreeBanks for the languages with a relatively simple morphology and the fixed word order. Data-driven treebanking is now at the state where naturally occurring text in the news domain can be automatically annotated with high accuracy according to standard parsing evaluation measures. However, when moving from languages with relatively fixed word order to languages with richer morphologies and less-rigid word orders, the standard issues for annotation TreeBanks developed for languages with fixed word order exhibit a large drop in accuracy.

3. Creating a Georgian Treebank and a Vanilla CFG

There are constituent TreeBanks for several languages in existence, along with a very limited number of parsing reports on them. The main challenge of constituent parsing for morphologically rich languages is in the handling of the huge number of word forms. According to the reports, the size of the preterminal set in the standard context-free grammar environment is crucial. If we use only the main part-of-speech (POS) tags as preterminals (as is the case with the strongly configurational languages), a considerable amount of information, encoded in the morphological description of the tokens, will be lost. Nevertheless, using the full morphological description as preterminal labels yields a set of over a thousand preterminals, resulting in data sparsity and performance problems (Szántó et al., 2014).

With this in mind, in order to manually construct the Georgian syntactically annotated trees, we had to perform the following text processing procedures:

- tokenization
- morphological analysis
- POS tagging and syntactic annotation.

Tokenization and morphological analysis were done by the Finite-State Transducer for Georgian (Kapanadze, 2010).

Before starting syntactic annotation procedures for the Georgian text, we made an overview of experience in building parallel TreeBanks for languages with different structures (Megyesi and Dahlqvist, 2007, Grimes at al., 2011, Rios at al., 2009, Samuelsson and Volk, 2005).

In a Quechua-Spanish parallel TreeBank, due to strong agglutinative features of the Quechua language, the monolingual Quechua TreeBank was annotated on morphemes rather than words. This allowed to link morpho-syntactic information precisely to its source. Besides, according to the authors, building phrase structure trees over Quechua sentences does not capture the characteristics of the language. Therefore, for its description a Role and Reference Grammar has been opted that allowed by using nodes, edges and secondary edges to represent the most important aspects of Role and Reference syntax for Quechua sentences (Rios at al., 2009).

Georgian is also an agglutinative language that uses for a wordform building both, suffixing and prefixing, though, there is no need to annotate the Georgian TreeBank on morphemes. Therefore, morphological analysis is one of the basic issues for agglutinating languages, since it provides useful clues for resolving syntactic ambiguity, and the parsing model should have a way of utilizing these hints. A lexicon-based parse engine has been oriented to capture the specifics of the Georgian morphology manifesting rich syntactic clues (among others the syntactic valency) encapsulated in the finite verb forms.

Syntactic annotation procedures were carried out manually using the *Synpathy* tool (Synpathy: Syntax Editor, 2006). It drew on an adapted version of the TIGER-XML encoding scheme (Brants and Hansen, 2002) that employs a SyntaxViewer developed for the TIGER-Research project (Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart). The POS tags pursue STTS (Stuttgart-Tübinger Tagset) guidelines with the necessary changes relevant to the Georgian grammar formal description and has been tested in the CLARIN-D project for the GRUG TreeBank repository building (Kapanadze, 2017).

In Georgian, as in many other languages, word order is much more flexible (for example, the subject may appear

either before or after a verb, etc.) as a result of its rich and productive morphology. In languages with flexible word

order the meaning of the sentence is realized using other structural elements, like word inflections or markers, which reflect morphological information.

A preferred basic word order without a Theme/Rheme bias for Georgian is SOV. The most notable feature in a syntactic description model for the Georgian clause is a phenomenon classified as a mutual government and agreement relations between verb-predicate and its actants (resp. NP), which number may reach up to three in a single clause. This anticipates control of the noun declension case markers by verbs, whereas, in its turn, the verb formants for person and number are governed by nouns presented in the clause. As a consequence of the verb-predicate capability to reflect morphologically the agreement relations with actants - Subject (SB), Direct Object (DO), Indirect Object (IO) as pronouns - can be omitted in the word order without a consequence for the clause meaning comprehension. The "reduced" clauses are equally "eligible" as their source ones in terms of the clause meaning representation.

In Figure 1 a syntactic tree of a Georgian complex sentence (*) as an outcome of the CFG parse procedure is depicted.

(*) თუ ღმერთი გწამთ, არ მითხრათ ახლა, რომ შავი თეთრია.

(Lit. "If you believe in god (=For god's sake), do not tell me now that black is white").

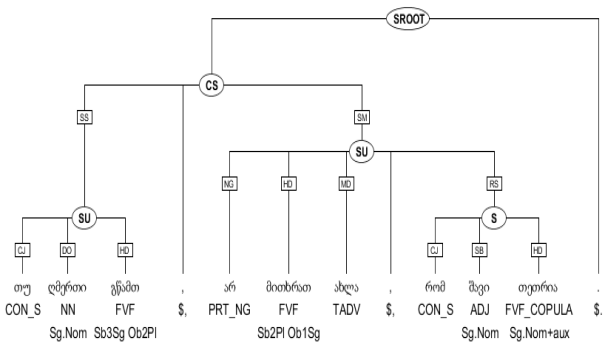


Figure 1: An adapted TIGER-XML scheme for a Georgian sentence.

The sentence in Figure 1 visualizes a hybrid approach to the syntactic annotation issue as the tree-like graphs and integrates annotation according to the constituency representations and functional relations. In a tree structure the node labels are phrasal categories.

The Complex Sentence (CS) in Figure 1 has two clauses as constituents:

– A clause without a Subject (SU) on the left side. As a daughter node it is linked by SS (*Subordinate Sentence*) Relation marked on an edge label.
– A simple sentence (S) on the right side. It is linked as a second daughter node by SM (*Main Sentence*) Relation that is marked on the consequent edge label.
– In its turn, S on the right side enjoys a simple sentence (S) as a daughter node linked by the RC (*Relational Construction*) as a constituent.
The edge labels for terminal nodes display the following syntactic functions: Conjunction (CJ), Subject (SB), Head (HD), Modifier (MD), Direct Object (DO).

The tokens in terminal nodes are annotated with POS tags such as Subordinating Conjunction (S_CNJ), Normal Noun (NN), Adjective (ADJ), Finite Verb (FVF), Negation Particle (N_PRT) and Adverb (ADV). They are saturated with morphological features of number (Sg) and case (Nom) for Normal Nouns. The Finite Verbs are annotated with features for person and number of subject and object (Sb3Sg Ob3Pl) (Sb2pl Ob1Sg), though, the Subject in the left constituent (SU) and both - the Subject and the Object in right (SU) one - are ommited in the word order. Thus, the monolingual treebanks converted into TIGER-XML format are a powerful database-oriented representation for graph structures in which each leaf (= token) and each node (= linguistic constituent) has a unique identifier.

Further, drawing on the sketched principle, we had manually built around 300 high quality morphologically and syntactically annotated trees. This repository had been used as training data for extracting a vanilla Context-Free Grammar and a lexicon for the Georgian language. The number of rules extracted from the syntactically annotated sentences has exceeded 1000. However, the rules are extracted with respect just to POS without morphological fetures as it is adopted in general while developing CFG parsers.

4. Conclusion and Future Plans

In the future we intend to implement a mixed syntactic parsing method for the Georgian text that will utilize a traditional CFG approach combined with a morphological feature commonly known as syntactic valency of a verb-predicate. Morphological information of valency value will be extracted from verb which normally is the head (HD) of a clause.

E.g. In Figure 1:

- FVF - გწამთ - in SU (the left-hand constituent)
- FVF - მითხრათ - in SU (the right-hand constituent)
- FVF_COPULA - თეთრია - in S (the secondary node in the right-hand constituent)

The verb syntactic valency feature will be used for determining syntactic structure of a clause in syntactic trees. To this end in the meantime we are developing a

new version of a Finite-State Morphoparser that will provide the Georgian verb parse output (alongside the POS tag) with the valency data.

For building a full-scale Georgian syntactic parser, we also intend to make use of the developed vanilla CFG that was extracted from the monolingual Georgian treebank. It will be utilized for finding optimal morphological features/preterminals for implementation in a Probabilistic Context-Free Grammar (PCFG) parser. The reason for such decision is the advantage of a deterministic part-of-speech tagger that can produce a morphologically annotated Georgian corpus achieving almost 100% accuracy after manual disambiguation (Kapanadze, 2010) and providing the tokens with POS saturated also with morphological information using features such as case, number for nouns and adjectives, and person, tense, syntactic valency for verbs.

In parallel we will extend a monolingual lexicon extracted from the Georgian TreeBank by adding all possible case forms for Nouns in singular and plural for each lexicon entry (14 forms for modern and 5 forms for old Georgian plural). For the mentioned procedures a Georgian FST morphological generator is intended to utilize.

According to the reports, the most successful supervised constituent parsers at the first stage apply a PCFG to extract possible parses. The *n*-best list parsers keep just the 50-100 best parses according to the PCFG. These feature templates exploit atomic morphological features and achieve improvements over the standard feature set. These methods use a large feature set — usually a few million features — and are engineered for English (Szántó and Farkas, 2014).

The innovative aspect of the proposed approach is a unique procedure for finding the optimal set of preterminals by merging morphological feature values.

The main advantage of this methodology over previous undertakings is the performance speed — it operates inside a PCFG instead of using a parser as a black box with retraining for every evaluation of a feature combination — and it can investigate particular morphological feature values instead of removing a feature with all of its values (Szántó and Farkas, 2014).

5. Bibliographical References

Fraser, A., Schmid, H., Farkas, R., Wang, R. and Schütze, H. (2013). Knowledge sources for constituent parsing of German, a morphologically rich and less-configurational language. In *Computational Linguistics*, Volume 39, Issue 1. MIT Press Cambridge, Ma, USA.

Kapanadze, O. (2010). Describing Georgian Morphology with a Finite-State System. In A. Yli-Jura et al. (Eds.): *Finite-State Methods and Natural Language Processing 2009, Lecture Notes in Artificial Intelligence*, Volume 6062, pp.114-122, Springer-Verlag, Berlin Heidelberg.

Szántó, Z. and Farkas, R. (2014). Special Techniques for Constituent Parsing of Morphologically Rich Languages. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Gothenburg, Sweden.

Synphaty: Syntax Editor. (2006). Manual – Nijmegen: Max Planck Institute for Psycholinguistics. The Netherlands.

Brants, S. and Hansen, S. (2000). Developments in the TIGER Annotation Scheme and their Realization in the Corpus. In *Proceedings of the Third Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, pp. 1643–1649.

Kapanadze, O. (2017). *Multilingual GRUG Parallel TreeBank — Ideas and Methods*. LAMBERT Academic Publisher. 52 p. ISBN-13: 978-3-330-34810-3. EAN: 9783330348103.

Megyesi, B. and Dahlqvist, B. (2007). A Turkish-Swedish Parallel Corpus and Tools for its Creation. In *Proceedings of Nordiska Datalingvistdagarna (NoDaL- iDa 2007)*.

Grimes, S, Li, X., Bies, A., Kulick, S., Ma, X. And Strassel, S. (2011). Creating Arabic-English Parallel Word-Aligned Treebank Corpora at LDC. In *Proceedings of the Second Workshop on Annotation and Exploitation of Parallel Corpora. The 8th International Conference on Recent Advances in Natural Language Processing (RANLP 2011)*. Hissar, Bulgaria.

Rios, A., Göhring, A. and Volk, M. (2009). Quechua-Spanish Parallel Treebank. In *7th Conference on Treebanks and Linguistic Theories*, Groningen. The Netherlands.

Megyesi, B., Hein Sågvall, A., Csató E.A. and Johanson, E. (2006). Building a Swedish-Turkish Parallel Corpus. In *Proceedings of the FifthInternational Conference on Language Resources and Evaluation (LREC 2006)*. Genoa. Italy.

A multilingual German – Russian – Ukrainian - Georgian Parallel Treebank. http://fedora.clarin-d.uni-saarland.de/grug/

Samuelsson, Y. and Volk, M. (2005). Presentation and Representation of Parallel Treebanks. In *Proceedings of the Treebank-Workshop at Nodalida*. Joensuu, Finland.

Proceedings, pages 346–349
Paris, UNESCO Headquarters, 5-6 December, 2019.

On Practical Realisation of Autosegmental Representations in Lexical Transducers of Tonal Bantu Languages

Anssi Yli-Jyrä
Department of Digital Humanities, University of Helsinki
PL 24, 00014, FINLAND
anssi.yli-jyra@helsinki.fi

Abstract

A lexical transducer is a language technology resource that is typically used to predict the orthographic word forms and to model the relation between the lexical and the surface word forms of a morphologically complex language. This paper motivates the construction of tone-enhanced lexical transducers for tonal languages and gives two supporting arguments for the feasibility of finite-state compilation of autosegmental derivations. According to the Common Timeline Argument, adding a common timeline to autosegmental representations is crucial for their computational processing. According to the Compilation Argument, the compilation of autosegmental grammars requires combining code-theoretic and model-theoretic research lines.

Keywords: lexical transducers, tonal languages, autosegmental phonology, compositionality, origin correspondence

Résumé (in Swahili)

Makala hii inahamasisha umuhimu wa transdyusa za kimsamiati zinazoainisha pia toni kwa lugha zenye toni, (kama vile lugha nyingi za Kibantu), na inadai kwamba sasa kuna uwezekano wa kuunda fomalismu ya finite-state kwa sarufi za autosegmentali na kwa transdyusa za kimsamiati zinazoainisha pia toni.

1. Motivation

Lexical transducers (LT) are finite state machines that are specialized to the description of the relation between the word forms and the corresponding morphological analyses. They constitute, for many languages, the main approach to implement the Morphological Component in the conservative BLARK scheme for language technology development. Thus, they are to be considered an essential resource for language technology development of any synthetic (fusional, agglutinative and polysynthetic) language.

In the era of statistical natural language processing and highly successful neural network models, it is relevant to ask whether we still need lexical transducers that are typically constructed manually by a linguist and a lexicographer. The answer is clear: *lexical transducers have particular strengths when there are very limited resources, such as annotated texts, available.* In this context, lexical transducers are very useful as a minimal adequate teacher, or in producing gold morpheme annotation to texts of agglutinative languages in order to train statistical models that learn to produce similar morpheme labels. For the same reason, they have been used to establish an orthography and to maintain the prescriptive spelling of languages that have a scarse written tradition.

Lexical transducers are *more compact, efficient, interpretable, testable and debuggable* than most statistical models. They can be used in checking and validating morphological and phonological grammas and statistical models. Besides this, a lexical transducer can serve as a distilled alternative for a functionally equivalent neural network that has a large memory footprint and whose behaviour needs to be tested and verified against linguistic parameters and new observations.

Tone (Yip, 2002) is a phonologically meaningful pitch distinction that complements vowels and consonants in some spoken languages, forming an autonomous string in the whole phonological representation of sentences. Up to seventy percent of the world's languages are tonal, i.e. they use tone to encode important lexical and grammatical distinctions. To describe the word forms of these languages precisely, one needs to model the relation between the lexical forms and the surface tone melodies. This relation is complex, especially when one considers morphologically rich tone languages in the Bantu family and the context dependent, morpho-syntactic tone alternations.

Because tone disambiguates the meaning between otherwise similar word forms, a *phonologically motivated orthography* explicates the tonal distinctions between similar word forms of a tonal language. This means that if we build a lexical tranducer for a tonal language, the computational and linguistic description of its word forms should include the meaningful tone distinctions in the surface word form representation. This is especially important if the morphology is complex and the tone is subject to complex phonological alternations that do not allow treating tone as a segmental feature of the lexical word forms.

A lexical transducer that is describing morphological and morpho-syntactic tone variation will be called *tone-enhanced*. Although tone is not always indicated in the adopted orthographies, tone-enhanced lexical transducers with tone markup are very valuable tools in the development of models that restore the morpho-syntactically determined tone melodies in texts that do not yet indicate tone melody contrasts.

Construction of a lexical transducer is typically a very practical effort. Traditionally, practical resource construction has been based on Two-Level Phonology and Morphology, Paradigmatic Morphology, or classical forms of Generative Phonology. More advanced theories such as Autosegmental Phonology, Optimality Theory, Correspondence Theory, Domain Theory, Q-Theory, and Harmonic Serialism do not yet facilitate a regular construction of lexical transducers.

Autosegmental Phonology (AP) (Goldsmith, 1979) is the first major extension of Generative Phonology towards tonal grammars, having such innovations as *morphemic tone* and the *autonomy* of the tonal tier. Today, AP is still one of the most useful and most widely understood phonological theories for the description of tonal alternations in field linguistics. Besides this, AP provides us an important multi-tiered phonological representation that forms a starting point for the development of more recent phonological representations and theories. It is, therefore, natural that we develop some AP-based lexical transducers before trying to construct lexical transducers that are based on more advanced theories.

2. The Prior Work

In Computational Morphology and Phonology, one of the most fundamental findings has been that phonological derivations correspond to finite-state relations. This result has led to the development of finite-state methods in natural language processing, including the method that constructs lexical transducers. Muhirwe (2010) treated tone as a segmental phenomenon in a lexical transducer. However, we have not yet been able to compile a large-scale, autosegmental tonal grammar and lexicon into an equivalent finite-state transducer – a tone-enhanced lexical transducer. This is due to three major AP-related computing challenges:

1. the storing of the autosegmental representations
2. the formation of the underlying representations
3. the input-output correspondences in transducers.

Previous research has aimed at *storing* autosegmental representations for algorithmic manipulation. The approaches include codes (Kornai, 1995; Yli-Jyrä, 2015; Jardine and Heinz, 2015; Yli-Jyrä, 2016), strings of tuples (Kiraz, 2000), multi-grained strings (van Leeuwen and te Lindert, 1991; Eisner, 1997; Yli-Jyrä and Niemi, 2006; Barthélemy, 2007; Yli-Jyrä, 2013), tuples of strings (Kay, 1987; Wiebe, 1992), and string sets (Bird and Ellison, 1994). Ideally, the encoding function should be a concatenation homomorphism, but Wiebe (1992) showed that no linear encoding satisfies this requirement when the tiers are not synchronised. However, there are restricted subsets of single-timeline autosegmental representations that are both closed under concatenation and homomorphically i.e. compositionally encodable (Yli-Jyrä, 2015; Jardine and Heinz, 2015; Yli-Jyrä, 2019a).

Computational *formation of underlying representations* (URs) has been addressed in Yli-Jyrä (2013) with a naive and deterministic association rule that works for a finite set of previously known morphemic melodies. Alternatively, the lexicon can consist of sequences of ready-made underlying representations of morphemes or local patterns.

The input-output correspondences and the expressive power of autosegmental grammars and their learning problem have been studied recently from the perspective of finite model theory and grammar inference (Jardine, 2014; Jardine, 2017a). Yli-Jyrä (2013) presented a practical transducer compilation method under certain restrictions on tone patterns. This method treated tone and associations naively as span markup in the segmental tier. This paper demonstrates a powerful method for compiling multi-component autosegmental rules over encoded grahs.

3. This Paper

Besides giving the motivation to build tone-enhanced lexical transducers, the current short work aims at arguing that the recently found string encoding for arbitrary single-timeline graphs (Yli-Jyrä, 2019b) opens a new technological opportunity. Investment in this opportunity extends the current finite-state technology for lexical transducer construction with the notion of rewritable and constrainable graph structure over a (discrete) timeline.[1]

The Common-Timeline Argument. Section 4. argues that (i) by storing the melody and the segments of underlying morphemes on a single timeline, we finally obtain an encoding for the graph structure of the underlying representation, and that (ii) the timeline can be maintained during the phonological processing.

The Compilation Argument. Section 5. argues that (i) a previously developed compilation method for a family of autosegmental rewriting rules can be generlized, but we need (ii) a higher-level logical formalism whose formal semantics links the specified rules into such low-level rules that we currently can compile.

4. The Common Timeline Argument

In the construction of lexical transducers, the described sets and relations over phonological representations need to have *good closure properties*. Especially *concatenation* and *Boolean operations* are extremely important.

Untamed autosegmental phonological representations are an impractical idea. They can be viewed naively as a logical description for an equivalence class of graphs or association drawings that differ with respect to the linear representation of the floating tone autosegments and unassociated segments. Such drawings and their classes are deficient wrt closure properties as their concatenation is not compositional (Wiebe, 1992). Morever, while we can use multi-tape finite-state machines

[1] The current work considers only finite graphs with linearly ordered nodes (=discrete bounded timeline) but is extensible to infinite graphs over a discrete timeline.

to recognize associationless autosegmental representations, such two-tape machines are not closed under Boolean operations as their emptyness is decidable but their equivalence is not (Griffiths, 1968).
Consequently, autosegmental representations need some form of internal *synchronisation* to be computationally well-behaving and closed under important operations. A natural way to introduce synchronisation is to assume morphemic (Yli-Jyrä, 2013) or otherwise local tone association patterns (Jardine and Heinz, 2015; Jardine, 2017b), or by specifiying the linearisation of unassociated elements (*inertial* autosegmental representations) (Yli-Jyrä, 2015). With synchronisation, the tiers of the underlying autosegmental representations can be interleaved to a *single timeline*. This common timeline for the underlying tiers should not be confused with the notion of the timing tier. Under a single timeline, autosegmental representations, viewed as graphs, have a string encoding that is bijective and respects concatenation (Yli-Jyrä, 2019b). This enables the description of regular subsets of the code strings and corresponding single-timeline graphs and gives these subsets their good closure properties.
The single underlying timeline fixes the origin of tones but allows the *independence* of tiers: floating means lack of association, and linking, shifting, spreading, reduplication and metathesis are processes that alter the association edges without affecting the timeline. The altered associations indicate the remapping of the timeline of the underlying tone tier to the timeline of the segmental tier, but both tiers originate from a shared underlying timeline.
The separation of the origin information from the associations has *advantages*: (i) the shared timeline encodes input-output correspondences. (ii) One may also combine the input and output autosegmental graphs of a phonological mapping and build a union graph that is represented in a single timeline and subject to well-formedness or faithfulness constraits. (iii) One may observe machine learnable local patterns in different stages of the phonological processing, in the origin structure, and in the correspondences.

5. The Compilation Argument

Lexical transducers of Bantu languages contain an infinite number of (compound) words and millions of inflections. Due to all these word forms in the transducer, it is not feasible to apply classical graph rewriting methods. Instead, the lexicon has to be constructed by *composition* of regular, possibly infinite relations that are recognized by finite-state transducers. With the single-timeline encoding, we can construct *constraints and conditions of phonological alternations*. In particular, we can

1. compile one-level autosegmental constraints and context conditions, including graph-based local constraints
2. describe feasible changes between input and output autosegmental representations
3. compile constraints on input-output correspondences.

The conditions of rules are the basis for their practical compilation. Yli-Jyrä (2013) presented a method for compiling conditions of autosegmental rewriting rules into finite-state transducers. This method is a combination of (a) optional parallel rewriting with two-level context conditions (aka two-level context restrictions) and (b) a comparison-based output optimisation that converts the optional replacements into obligatory ones.
A tone can have an unbounded number of associations. Multiple association can correspond to edges in a graph (Yli-Jyrä, 2019b) or spans in the timeline (Yli-Jyrä, 2013). Using edges makes the specification of a speading rule is more complex than the specification where the span boundary is simply moved. Therefore, with the new encoding, a *high-level formalism* for autosegmental rewriting rules is necessary.
It is pretty likely that in real languages, autosegmental representations of their phonology can be embedded to a regular subset of code strings. Regular subsets of this space are closed under all important operations that are needed to define, for these relations, a specification formalism (such as Monadic Second-Order Logic over restricted autosegmental representations) that is mechanically compiled into finite-state transducers. This is the point where the encoding research (Yli-Jyrä, 2019b) can be expected meet research on phonological model-theory and learnability (Jardine, 2017a).

6. Conclusion

This article has explained why tone-enhanced lexical transducers are needed and why they can be now be constructed, using a combination of the code-theoretic and model-theoretic perspectives to the AP research. In particular, the article presents two supporting arguments for the feasibility of practical implementation of autosegmental derivations when constructing lexical transducers.

1. The Common Timeline Argument of this article states that it is computationally beneficial to assume that there is an underlying timeline that represents the temporal origin of the tones and segments. Having an underlying timeline that is shared through the derivation steps is consistent with the idea of having independent tiers whose associations can be missing and do not need to respect accurately any underlying timeline.
2. The Compilation Argument of this article states that the existing knowlegde on rule compilation and the specification formalisms can be applied in order to develop a practical formalism for two-tiered AP.

The author is looking forward to possibilities to implement practical tone-enhanced lexical transducers for tonal languages, especially for morphologically complex ones such as found among the Bantu family that contains some 500 languages.

7. Acknowledgements

Writing this article has been enabled by the mobility grant by the University of Helsinki (Faculty of Arts N2/2017), allowing to visit the Hebrew University. The research has had synergy with more general research on string encoding for syntactic and semantic graphs in Research Fellowship projects (279354/273457/313478), and with research on manually aligned parallel texts as a Faculty supported University Researcher (2019). The author started inquiries into the computational implementation of Bantu tone in 2010 under funding of the Academy of Finland, via a Development Research project (2010: 134614 – *MDGs and African language technology: Roadmap to the development of Bantu language resources*). The author is grateful to O. Abend, E. Kuriyozov, J. Tiedemann and C. Gómez Rodríguez and F. Drewes for inspiring discussions, and A. Jardine, A. Fleisch, D. Killian, A. Kornai, and L. Aunio for help with tonal phonology, and A. Hurskainen for the Bantu tone challenge in 2007 and help with the Swahili translation.

8. Bibliographical References

Barthélemy, F. (2007). Multi-grain relations. In *Proceedings of the 12th International Conference on Implementation and Application of Automata*, pages 243–252, Berlin, Heidelberg. Springer-Verlag.

Bird, S. and Ellison, T. M. (1994). One-level phonology: autosegmental representations and rules as finite automata. *Computational Linguistics*, 20(1):55–90.

Eisner, J. (1997). Efficient generation in primitive optimality theory. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 313–320, Madrid, Spain, July. Association for Computational Linguistics.

Goldsmith, J. (1979). The aims of autosegmental phonology. In D. A. Dinnsen, editor, *Current approaches to phonological theory*, chapter 8, pages 202–222. Indiana University Press, Bloomington.

Griffiths, T. V. (1968). The unsolvability of the equivalence problem for Λ-free nondeterministic generalized machines. *J. ACM*, 15(3):409–413, July.

Jardine, A. and Heinz, J. (2015). A concatenation operation to derive autosegmental graphs. In *Proceedings of the 14th Meeting on the Mathematics of Language (MoL 2015)*, pages 139–151, Chicago, USA, July. Association for Computational Linguistics.

Jardine, A. (2014). Logic and the generative power of autosegmental phonology. In *Proceedings of the 2013 Annual Meeting of Phonology*.

Jardine, A. (2017a). The expressivity of autosegmental grammars. Manuscript, June.

Jardine, A. (2017b). The local nature of tone-association patterns. *Phonology*, 34(2):363–384.

Kay, M. (1987). Nonconcatenative finite-state morphology. In Bente Maegaard, editor, *3rd Conference of the European Chapter of the Association for Computational Linguistics*, pages 2–10. The Association for Computer Linguistics.

Kiraz, G. A. (2000). Multitiered nonlinear morphology using multitape finite automata: A case study on Syriac and Arabic. *Computational Linguistics*, 26(1):77–105, March.

Kornai, A. (1995). *Formal Phonology*. Garland Publishing, New York.

Muhirwe, J. (2010). Morphological analysis of tone marked Kinyarwanda text. In *Finite-State Methods and Natural Language Processing*, volume 6062 of *Lecture Notes in Computer Science*, pages 48–55. Springer Berlin Heidelberg.

van Leeuwen, H. and te Lindert, E. (1991). Speech maker: text-to-speech synthesis based on a multilevel, synchronized data structure. In *International Conference on Acoustics, Speech, and Signal Processing, 1991*, pages 781–784 vol.2, Apr.

Wiebe, B. (1992). Modelling autosegmental phonology with multitape finite state transducers. Master's thesis, Simon Fraser University.

Yip, M. (2002). *Tone*. Cambridge Studies in Linguistics. Cambridge University Press.

Yli-Jyrä, A. and Niemi, J. (2006). Pivotal synchronization languages: A framework for alignments. In Anssi Yli-Jyrä, et al., editors, *Finite-State Methods and Natural Language Processing*, volume 4002 of *Lecture Notes in Computer Science*, pages 271–282. Springer Berlin Heidelberg.

Yli-Jyrä, A. (2013). On finite-state tonology with autosegmental representations. In Mark-Jan Nederhof, editor, *Proceedings of the 11th International Conference on Finite State Methods and Natural Language Processing, St. Andrews, Scotland, UK, July 15-17, 2013*, pages 90–98. The Association for Computer Linguistics.

Yli-Jyrä, A. (2015). Three equivalent codes for autosegmental representations. In Thomas Hanneforth et al., editors, *Proceedings of the 12th International Conference on Finite-State Methods and Natural Language Processing, Düsseldorf, Germany, June 22-24, 2015*. The Association for Computer Linguistics.

Yli-Jyrä, A. (2016). Aligned multistring languages. In *TTATT 2016, Proceedings of the Workshop, Workshop on Trends in Tree Automata and Tree Transducers*, Seoul, South Korea, July 18.

Yli-Jyrä, A. (2019a). *Optimal Kornai-Karttunen Codes for Restricted Autosegmental Representations*. Number 224 in CSLI Lecture Notes. CSLI Publications, Stanford, USA.

Yli-Jyrä, A. (2019b). Transition-based coding and formal language theory for ordered digraphs. In *Proceedings of the 14th International Conference on Finite-State Methods and Natural Language Processing*, pages 118–131, Dresden, Germany, September. Association for Computational Linguistics.

Proceedings, pages 350–354
Paris, UNESCO Headquarters, 5-6 December, 2019.

Text-Independent Dialect Classification in Read and Spontaneous Speech

Oliver Jokisch and Johanna Dobbriner
Leipzig University of Telecommunications (HfTL), Institute of Communications Engineering, Germany
Technological University Dublin, School of Computing, Ireland
jokisch@hft-leipzig.de, johanna.dobbriner@gmail.com

Abstract

Linguistic diversity and the fundamental freedom of users of language technology (LT) to access information and knowledge in their own language(s) or dialect(s) lead to certain requirements with regard to truly multilingual language technologies, in particular for under-resourced languages and application domains. One key issue is the low-threshold creation of high-quality speech corpora and the corresponding annotation data to train powerful analysis or classification algorithms as a base for state-of-the-art language technology. Dialects constitute an important part of the mentioned linguistic diversity. In this contribution, we shortly discuss basic concepts of automatic dialect classification (ADC) with a focus on methods that do not require expensive prior annotation or labeling. Starting with text-independent ADC methods for well-studied major languages and summarizing our results of a case study on read and spontaneous German, we convey necessary development steps for under-resourced language data and a possible processing chain.
Keywords: dialect classification, read and spontaneous speech, under-resourced language, corpus design, feature selection

Résumé

Sprachliche Vielfalt und ein Grundrecht von Sprachtechnologie-Anwendern, auf Informationen und Wissen in ihrer eigenen Sprache oder sogar ihrem eigenen Dialekt zuzugreifen, führen zu dezidierten Anforderungen an tatsächlich mehrsprachige Technologien, insbesondere für Sprachen und Anwendungsdomänen mit geringen Ressourcen. Das zentrale Thema ist eine niedrigschwellige Erstellung hochwertiger Sprachkorpora und zugehöriger Annotationsdaten, um leistungsstarke Analyse- oder Klassifizierungsalgorithmen zu trainieren, die eine Grundlage aktueller Sprachtechnologie darstellen. Dialekte sind ein wesentlicher Bestandteil der genannten sprachlichen Vielfalt. In diesem Beitrag diskutieren wir Konzepte der automatischen Dialektklassifizierung (ADC) mit dem Fokus auf Methoden, die keine aufwendige, vorherige Annotation erfordern. Ausgehend von textunabhängigen ADC-Methoden für ausgiebig untersuchte Hauptsprachen und einer Zusammenfassung der Ergebnisse einer Fallstudie zu gelesenem und spontanem Deutsch leiten wir Entwicklungs- und Verarbeitungsschritte für unterrepräsentierte Sprachdaten ab.

1. Dialect Classification in Major Languages

Particularly from a perspective of language and speech technologies, the differences between dialects are commonly less distinctive than the ones between single languages or language groups. Typical intra-language variations are small, with less-defined borders between dialect realizations. Both native and migrant speakers often exhibit a mixture of different dialects with regard to their vita. Therefore, Automatic Dialect Classification (ADC) combines known principles from language and speaker identification to automatically recognize a regional dialect of a given language from speech samples or corresponding transcripts. ADC methods can constitute a complementary technology in the area of under-resourced languages, e.g. to increase the performance of Automatic-Speech-Recognition (ASR) modules or to enable Intelligent-Language-Tutoring (ILT) systems with the goal to reduce or even to improve a regional accent.

Methods of accent reduction or improvement require robust ADC algorithms to identify the dialect and to evaluate the training progress, preferably avoiding transcribed speech. The so-called text-independent dialect classification has been researched by several authors, e.g. for English (Hanani et al., 2013; Najafian et al., 2018; Wang and van Heuven, 2018; Brown, 2016), Arabic (Bougrine et al., 2017; Biadsy et al., 2009; Akbacak et al., 2011) and Chinese (Zheng et al., 2005; Hou et al., 2010; Lei and Hansen, 2011). The ADC approaches can be roughly categorized as either acoustic/phonetic (Torres-Carrasquillo et al., 2008; Biadsy et al., 2010; Biadsy, 2011), phonotactic (Biadsy et al., 2009; Akbacak et al., 2011; Zissman et al., 1996) or prosodic (Bougrine et al., 2017; Chittaragi et al., 2017) including variations and combinations in features, modeling and classification methods, cf. (Najafian et al., 2016; Zhang et al., 2013). The most common ADC approaches rely on Mel-Frequency Cepstral Coefficients (MFCCs) for feature analysis and Gaussian Mixture Models (GMMs) with a Universal Background Model (UBM) for the classification task, followed by UBM adaptation to each of the target dialects (Hanani et al., 2013; Brown, 2016; Liu and Hansen, 2011; Lazaridis et al., 2014).

Apart from the above-mentioned works on major languages, text-independent dialect classification is still an under-researched and under-resourced topic – even with regard to German as a major European language. However, German dialect classification, based on phonotactic and acoustic approaches, was previously studied as part of an ASR system for broadcast speech (Stadtschnitzer, 2018).

Following the ADC approaches in other major languages we summarize our results of previous case studies on German ADC (Dobbriner and Jokisch, 2019a; Dobbriner and Jokisch, 2019b) with restricted training data, in which we tested various feature combinations for read and spontaneous speech from two corpora with 500 and 830 speakers respectively. Both modes of speech differ in multiple ways, so they are not pooled in the same model. Afterwards, we discuss the lessons learned from the viewpoint of under-resourced language data.

2. Text-Independent Dialect Classification in Read and Spontaneous German Speech

2.1. Speech Corpora

There are various German speech databases for multiple tasks within LT research and development, including selected databases with regionally accented speech, such as "Regional Variants of German 1" (RVG) (Burger and Schiel, 1998) and "Deutsch Heute" (DH) (Kleiner, 2015), that we used in our study (Dobbriner and Jokisch, 2019a). Both corpora are well-annotated and appropriate for many linguistic studies. In terms of training and test material, in particular for state-of-the-art methods in (deep) learning, a few hundred speakers with a few ten phrases per speaker has to be treated as low-resourced data.

RVG is a corpus within the BAS CLARIN Repository (Burger and Schiel, 1998), which comprises recordings of 500 speakers from nine different dialect regions in Germany. There are samples of about 1 min. of spontaneous speech as well as single numbers, commands and 30 phrases per speaker, recorded by four microphones simultaneously. The corpus is divided into nine dialect regions, illustrated by the sample speakers in Figure 1 with regard to their current home when the corpus was recorded, 1996 – 1997.

The DH corpus, recorded 2006 – 2009 in Germany, Switzerland and Austria by the Institute of the German Language (IDS) in Mannheim, contains variations in contemporary spoken German. In total, DH includes 830 speakers, mainly from high schools and further education centers. There are different parts of read and task-prompted speech. Contrary to RVG, the speakers in DH were not assigned a dialect in the database, i. e., we processed our own assignment with regard to (Mettke, 1989) and (Burger and Schiel, 1998). Table 1 summarizes the number of speakers in the different regions and dialects. The RVG speakers are not distributed uniformly over Germany, and the dialect regions themselves vary in size, which leads to imbalanced classes. The DH recording sites, on the other hand, are uniformly distributed, but the varying size of the dialect regions leads to imbalanced classes as well.

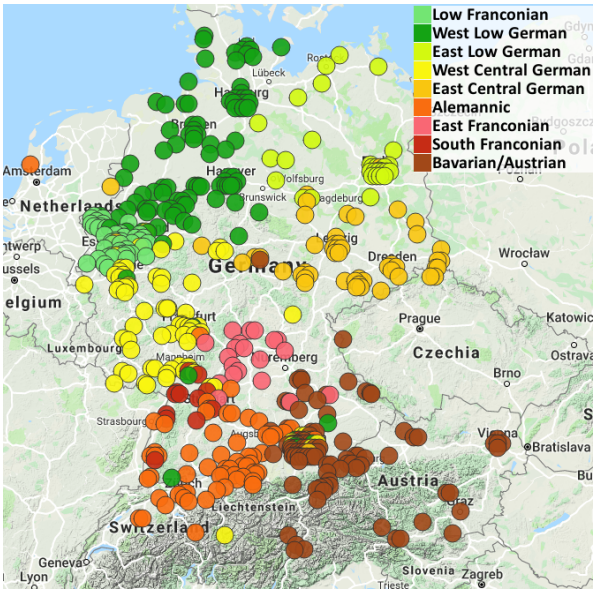


Figure 1: Dialects of RVG speakers by current home

Table 1: No. of speakers per dialect and spontaneous/read subcorpora, extension S/R (Dobbriner and Jokisch, 2019a)

Region	Dialect	Speakers RVG-S	RVG-R	DH-R
North	A Low Franconian	44	44	20
	B West Low German	103	103	149
	C East Low German	31	31	67
Center	D West Central German	73	73	128
	E East Central German	52	53	76
South	F Alemannic / Swabian	63	63	145
	G East Franconian	19	20	39
	H South Franconian	10	10	26
	I Bavarian / Austrian	100	100	179

2.2. Parameter Extraction and Classification

To compare ADC for spoken versus read German, we developed a tool chain (Dobbriner and Jokisch, 2019b) according to the GMM-UBM approach, which is comprised of the following steps:

1. Feature extraction
2. Feature processing
3. Computing the UBM
4. UBM adaptation to different dialects
5. Scoring of test samples for each dialect model
6. Classification test.

The first step consists of extracting Mel-Frequency Cepstral Coefficients (MFCC) with a sampling rate of 8kHz, frame length of 25ms, Hamming-windowing and 10ms frame shift. The resulting feature vectors consist of 12 MFCC and the spectral energy per frame. The feature vectors are processed in Step 2 by using Voice Activity Detection (VAD) through an energy threshold, RASTA-filtering and Cepstral Mean and Variance Normalization (CMVN). Additionally, delta and double delta, and Shifted Delta Cepstra (SDC) are computed from the MFCC to incorporate temporal context for each frame. In later experiments a sampling rate of 16kHz with similar overall results (Dobbriner, 2019) was tested, which led to higher calculation complexity. Afterwards, the speech data is randomly divided into a speaker-disjunct training set and a test set. Step 3 is accumulating the feature vectors of all training speakers and training the UBM by Expectation-Maximization (EM) for 256 or 512 gaussians, which had proven to be successful in prior ADC research. In step 4, the maximum-a-posteriori (MAP) algorithm is used, to adapt the means of the UBM to each dialect by using all speakers of this dialect category in the training set. MAP adapts the measure of interest (in our case the means of each gaussian in the UBM) until the probability of all data is maximized in the distributions of the

adapted model. All test samples are scored in step 5 for every adapted model using log-likelihood, and the highest score per sample is determined as the corresponding dialect. Lastly, the weighted accuracy of the model is calculated by dividing all correctly classified test samples per class by the total number of test samples per class, aiming at the average accuracy over all classes. We always refer to a weighted accuracy measure, since our classes are imbalanced in their number of speakers, cf. Table 1.

Our ADC processing chain, including feature extraction, classification and evaluation is based on the Python toolkit "Sidekit" (Larcher et al., 2016), which was originally designed for speaker identification with a certain similarity to the task of dialect classification.

2.3. Experiments and Results

To realistically experiment with our restricted German dialect data, we switched between a coarse-grained dialect classification, which divided the speakers into just three main regions (low/North, central and upper/South German), that are widely agreed among linguists, and a fine-grained partition of both corpora, RVG and DH, into nine dialect regions with the disadvantage of sparse training data in some classes. Longer speech samples and monologues up to one minute proved to be suitable for training and testing of our classifiers on spontaneous speech in "RVG-S" (Dobbriner and Jokisch, 2019b). While maintaining an appropriate sample duration for the classifiers and to increase the number of samples per dialect, we therefore concatenated approximately 30 read phrases per speaker in "RVG-R" to three files of ten phrases each. For subcorpus "DH-R", we selected one minute of read speech per speaker for training and testing the ADC system.

All speakers were randomly partitioned into speaker-disjunct sets for training (80%) and testing (20%). Applying the processing chain described in section 2.2., we systematically analyzed different feature combinations in the three- and nine-dialect classification, based on RASTA, Delta/DoubleDelta, SDC and CMVN. Beside the best method from our previous study (GMM based on 512 gaussians), we tested the same feature combinations with 256 gaussians only to determine, whether similar accuracies can be achieved by smaller models with lower calculation complexity. Furthermore, to visualize effects of the random speaker selection, we repeated the classification. The different train/test-set constellations are marked by "0" or "1" in the following. As an example for spontaneous speech (RVG-S corpus), Figure 2 summarizes the test results of the more challenging nine-dialect classification: The weighted accuracies are sorted by feature combination and model type. Colored bars represent the spontaneous/read subcorpora (RVG-S, RVG-R, DH-R) and concrete test sets (0/1).

Considering a chance level of 11.1 %, the nine-dialect classification in the test set stretches over a huge range from a weighted overall accuracy of 14.0% to 35.3%, and the two test sets per corpus behave diverse too. In a three-dialect classification, the overall accuracies span from 37.6%, which is barely above chance level (33.3%), up to 56.0%. In both classification scenarios, the RVG-R subcorpus reached the highest accuracies, while the test results on

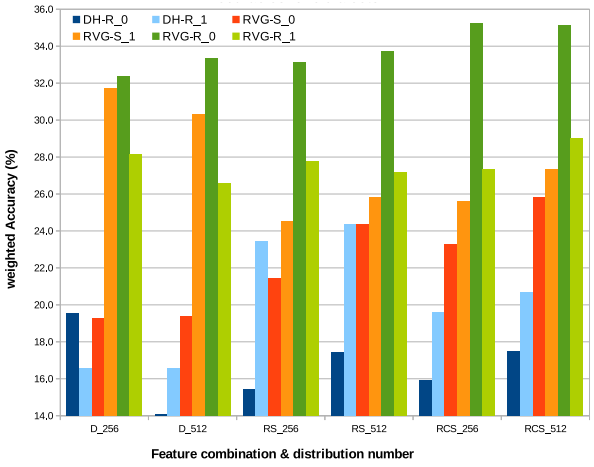


Figure 2: Nine-dialect classification for varying feature sets and 256/512 GMMs: R – RASTA, D – Delta/DoubleDelta, S – SDC, C – CMVN (Dobbriner and Jokisch, 2019a)

DH-R were significantly worse. For the RVG-S subcorpus, the RCS feature combination performed best, whereas the highest accuracies for DH-R were reached with the RS feature set. More gaussians in the GMM lead to higher classification accuracies – in (Dobbriner and Jokisch, 2019b) we surveyed constellations of 64 ...512 gaussians. Overall, the resulting accuracies are far from optimal but certainly above chance level. As a baseline system, the GMM-UBM approach seems to be effective for distinguishing dialect samples, and aside from the model quality, there are some potential reasons for the low accuracies observed, e.g. some dialects and dialect groups are similar and therefore hard to distinguish, even for human listeners. Besides, a few speakers in the test corpora articulate close to standard German, which may contradict to the forced-assigned dialects in the training. A general shortcoming is the unequal distribution of speakers per dialect as well as the unbalanced size of dialect regions. For a three-dialect categorization, the variation of speech within the regions seems too high, so that our models are not specific enough to allow for a robust classification. As shown in the confusion matrix of the best three-dialect model in Table 2, both northern and southern German are distinguished relatively well, but the central region is frequently confused with the northern region.

The results of the nine-dialect classification on RVG-S are presented in the corresponding confusion matrix in Table 3.

Based on the described acoustic approach, our results for RVG-S – a weighted overall accuracy of 31.7% for nine dialects (and 53.2% for three dialects) – outperform the re-

Table 2: Three-dialect accuracies (RVG-S)

	North	Central	South
North	23	10	6
Central	9	10	10
South	5	6	23
Σ	37	26	39
Accuracy(%)	62.2	38.5	59.0

Table 3: Nine-dialect accuracies (RVG-S)

	A	B	C	D	E	F	G	H	I
A	7	1	0	6	1	1	1	0	2
B	0	4	0	4	3	1	0	0	3
C	0	3	4	1	5	2	0	1	1
D	0	5	0	2	0	3	2	0	3
E	0	1	1	0	0	1	0	0	0
F	2	2	1	0	1	3	0	0	1
G	0	3	0	0	1	0	0	0	0
H	0	1	0	1	0	0	1	1	1
I	0	1	1	1	0	2	0	0	9
Σ	9	21	7	15	11	13	4	2	20
Acc(%)	77.8	19.0	57.1	13.3	0.0	23.1	0.0	50.0	45.0

sults of a phonotactic approach on RVG-S (Stadtschnitzer, 2018), which led to a nine-dialect accuracy of 19.2% only. In contrast, a second approach in (Stadtschnitzer, 2018) with acoustic-spectral features and a convolutional neural network (CNN) classifier on a small, well-annotated corpus achieved 56.7% accuracy on four dialect classes and 77.1% on two classes, but due to the different corpus and other constellations of classes and speakers, a direct comparison with our results is not possible.

Some peculiarities of the German dialects and spontaneous vs. read speech as well as potential explanations for the observed confusions in RVG-S, RVG-R and DH-R are addressed in (Dobbriner, 2019). After some modifications in feature processing, and in particular by a manual correction and re-assignment of dialect speakers from the original RVG and DH corpora into other/partly new classes, our classification results could be improved. The nine-dialect accuracy could be thereby increased up to 36.3%. Another trial, only differentiating between standard and dialect speech, achieved at a maximum accuracy of 76.0%.

In general, the ADC accuracies of our basic GMM-UBM classification system are similar for spontaneous and read speech, which indicates that the distinguishable features of a speakers' dialect are based on the same mechanisms and less influenced by the speaking mode. Of course, the surveyed approach requires a sophisticated back-end classification method – in a further step we tested different classifiers like support vector machines, logistic regression or artificial neural networks but the accuracies diversified about 2% only for our example of a nine-dialect classification.

3. ADC Conception for Under-resourced Languages and some Conclusions

As introduced for major languages in section 1., our modeling via GMM-UBM has proved a good baseline performance for German dialect classification too, in particular in the context of the rather low amount of training data. The ADC task does not require transcribed speech. The proposed tool chain in section 2.2. – feature extraction and processing, UBM computing, maximum-a-posteriori (MAP)adaptation to different dialects and scoring of test samples for each dialect model – based on the Python toolkit "Sidekit", is appropriate for all relevant tasks in analysis and classification. Toolboxes like WEKA (Frank et al., 2016), can support the classification by alternative methods but the potential accuracy improvements seem to be limited. With regard to a few hundred speakers and phrases, deep learning techniques seem superfluous.

Acoustic, namely spectral, features like MFCC are suitable for the ADC task, and modifications in the feature analysis and processing offer potential for optimization. Our results suggest that ADC may also work on shorter audio samples below a length of 1 min. Phonotactic and prosodic measures can be applied as well, but they have been barely discriminative in our dialect classification tests. A combination of approaches such as the Phone-Supervector method in (Biadsy, 2011), combining conventional phone recognition and GMM-mean super vectors on Arabic ADC, is an interesting option, if adequate components like a phone recognizer are available for the language in question.

For under-resourced languages and applications, the design and annotation quality of the training corpus are the most significant factor of influence. Naturally, ADC training requires well-annotated classes that reflect current regional varieties of the language and contain a sufficiently large number of speakers as well as somewhat balanced classes. Regional varieties may even change within a decade due to our dynamic life environment including migration and media influence, although that may be less of a concern for languages spoken only in an isolated location. To cover more than two (standard vs. dialect speech) or three regional dialects, the minimum requirement is a few hundred speakers with a few tens of longer phrases. A corpus size of about 500+ speakers as in our experiments demands a precise annotation and effort in manual corrections.

To construct mid-size corpora up to a few thousand speakers, existing speech data from different sources with similar recording conditions can be merged, which usually calls for a new, consistent annotation of the samples according to dialect and strength of dialect, preferably by semi-automatic means. Local broadcast programs as in (Stadtschnitzer, 2018) can be an appropriate source of information and should be a good option for cooperation. Another, potentially low-cost method, is based on crowd-sourcing approaches to reach volunteers more easily and widespread, as demonstrated with the "Voice Äpp" for Swiss German (Leemann et al., 2015) and the "English Dialects App" for British varieties (Leemann et al., 2018).

4. Acknowledgment

We would like to thank IDS Mannheim for providing the corpus "Deutsch Heute" (partly used in our experiments). The open-source corpus "Regional Variants of Contemporary German" from the Bavarian Archive for Speech Signals/CLARIN Repository was also quite helpful.

5. Bibliographical References

Akbacak, M., Vergyri, D., Stolcke, A., Scheffer, N., and Mandal, A. (2011). Effective Arabic dialect classification using diverse phonotactic models. In *INTERSPEECH, Florence, Italy, August 2011*, pages 737–740.

Biadsy, F., Hirschberg, J., and Habash, N. (2009). Spoken Arabic dialect identification using phonotactic modeling. In *Proc. Workshop on Computational Approaches*

to Semitic Languages, SEMITIC@EACL 2009, Athens, March 2009, pages 53–61.

Biadsy, F., Hirschberg, J., and Collins, M. (2010). Dialect recognition using a phone-GMM-supervector-based SVM kernel. In *INTERSPEECH Makuhari, Japan, September 2010*, pages 753–756.

Biadsy, F. (2011). *Automatic Dialect and Accent Recognition and its Application to Speech Recognition*. Ph.D. thesis, Columbia University.

Bougrine, S., Cherroun, H., and Ziadi, D. (2017). Hierarchical classification for spoken Arabic dialect identification using prosody: Case of Algerian dialects. *CoRR*, abs/1703.10065.

Brown, G. (2016). Automatic accent recognition systems and the effects of data on performance. In *Odyssey: The Speaker and Language Recognition Workshop, Bilbao, Spain, June 2016*, pages 94–100.

Burger, S. and Schiel, F. (1998). RVG 1 - a database for regional variants of contemporary German. In *Proc. of the 1st Int. Conf. on Language Resources and Evaluation*, pages 1083–1087, Granada, Spain.

Chittaragi, N. B., Prakash, A., and Koolagudi, S. G. (2017). Dialect identification using spectral and prosodic features on single and ensemble classifiers. *Arabian Journal for Science and Engineering*, 43.

Dobbriner, J. and Jokisch, O. (2019a). Implementing and evaluating methods of dialect classification on read and spontaneous german speech. In *Proc. 8th ISCA Workshop on Speech and Language Technology in Education (SLaTE), September 2019*, pages 53–58, Graz, Austria.

Dobbriner, J. and Jokisch, O. (2019b). Towards a dialect classification in German speech samples. In *Proc. 21th Intern. Conf. Speech and Computer (SPECOM), August 2019*, pages 64–74, Istanbul, Turkey. Springer LNAI.

Dobbriner, J. (2019). Automatic Dialect Classification in Spoken German. Master's thesis, Univ. Leipzig/HfTL.

Frank, E., Hall, M. A., and Witten, I. H. (2016). *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", 4th Ed.* Morgan Kaufmann, Amsterdam.

Hanani, A., Russell, M. J., and Carey, M. J. (2013). Human and computer recognition of regional accents and ethnic groups from British English speech. *Computer Speech & Language*, 27:59–74.

Hou, J., Liu, Y., Zheng, T. F., Olsen, J. Ø., and Tian, J. (2010). Multi-layered features with SVM for Chinese accent identification. In *Intern. Conf. on Audio, Language and Image Processing*, pages 25–30.

Kleiner, S. (2015). 'Deutsch heute' und der Atlas zur Aussprache des deutschen Gebrauchsstandards. In *Regionale Variation des Deutschen*, pages 489–518. de Gruyter, Berlin/Boston.

Larcher, A., Lee, K. A., and Meignier, S. (2016). An extensible speaker identification sidekit in Python. In *IEEE Intern. Conf. on Acoustics, Speech and Signal Processing, ICASSP, Shanghai, March 2016*, pages 5095–5099.

Lazaridis, A., el Khoury, E., Goldman, J., Avanzi, M., Marcel, S., and Garner, P. N. (2014). Swiss french regional accent identification. In *Odyssey 2014: The Speaker and Language Recognition Workshop, Joensuu, Finland, June 16-19, 2014*.

Leemann, A., Kolly, M.-J., Goldman, J.-P., Dellwo, V., Hove, I., Almajai, I., Grimm, S., Robert, S., and Wanitsch, D. (2015). Voice äpp: a mobile app for crowdsourcing Swiss German dialect data. In *INTERSPEECH, Dresden, Germany, September 2015*, pages 2804–2808.

Leemann, A., Kolly, M.-J., and Britain, D. (2018). The English dialects app: The creation of a crowdsourced dialect corpus. *Ampersand*, 5:1-17.

Lei, Y. and Hansen, J. H. L. (2011). Dialect classification via text-independent training and testing for Arabic, Spanish, and Chinese. *IEEE Trans. Audio, Speech & Language Processing*, 19:85–96.

Liu, G. and Hansen, J. H. L. (2011). A systematic strategy for robust automatic dialect identification. In *Proc. 19th European Signal Processing Conference, EUSIPCO, Barcelona, August 2011*, pages 2138–2141.

Mettke, H. (1989). *Mittelhochdeutsche Grammatik*. Bibliographisches Institut, Leipzig, Germany.

Najafian, M., Safavi, S., Weber, P., and Russell, M. J. (2016). Identification of British English regional accents using fusion of i-vector and multi-accent phonotactic systems. In *Odyssey: The Speaker and Language Recognition Workshop, Bilbao, June 2016*, pages 132–139.

Najafian, M., Khurana, S., Shon, S., Ali, A., and Glass, J. R. (2018). Exploiting convolutional neural networks for phonotactic based dialect identification. In *IEEE Intern. Conf. on Acoustics, Speech and Signal Processing, ICASSP, Calgary, April 2018*, pages 5174–5178.

Stadtschnitzer, M. (2018). *Robust Speech Recognition for German and Dialectal Broadcast Programmes*. Ph.D. thesis, University of Bonn, Germany.

Torres-Carrasquillo, P. A., Sturim, D. E., Reynolds, D. A., and McCree, A. (2008). Eigen-channel compensation and discriminatively trained Gaussian mixture models for dialect and accent recognition. In *INTERSPEECH 2008, Brisbane, September 2008*, pages 723–726.

Wang, H. and van Heuven, V. J. (2018). Relative contribution of vowel quality and duration to native language identification in foreign-accented English. In *Proc. 2nd Intern. Conf. on Cryptography, Security and Privacy, ICCSP 2018, Guiyang, March 2018*, pages 16–20.

Zhang, Q., Boril, H., and Hansen, J. H. L. (2013). Supervector pre-processing for PRSVM-based Chinese and Arabic dialect identification. In *IEEE Intern. Conf. on Acoustics, Speech and Signal Processing, ICASSP, Vancouver, May 2013*, pages 7363–7367.

Zheng, Y., Sproat, R., Gu, L., Shafran, I., Zhou, H., Su, Y., Jurafsky, D., Starr, R., and Yoon, S.-Y. (2005). Accent detection and speech recognition for Shanghai-accented Mandarin. In *INTERSPEECH, Lisbon, Portugal, September 2005*, pages 217–220.

Zissman, M. A., Gleason, T. P., Rekart, D., and Losiewicz, B. L. (1996). Automatic dialect identification of extemporaneous conversational, Latin American Spanish speech. In *IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing Conference Proceedings, ICASSP, Atlanta, USA, May 1996*, pages 777–780.

Proceedings, pages 355–357
Paris, UNESCO Headquarters, 5-6 December, 2019.

Rediscovering Past Narrations: the Oral History of the Romanian Language Preserved Within the National Phonogramic Archive[1]

Oana Niculescu, Maria Marin, Daniela Răuțu
Institutul de Lingvistică al Academiei Române „Iorgu Iordan - Alexandru Rosetti”
Calea 13 Septembrie 13, București 050711
{oeniculescu, dialectologie, danarautu}@yahoo.com

Abstract

In this presentation we aim to deliver a key message related to the preservation of the Romanian National Phonogramic Archive (AFLR). The data gathered within the Archive (the richest, most inclusive and diversified collection of dialectal texts and ethno-linguistic recordings in Romania) are of immeasurable documentary value. Through the digitization and preservation of AFLR we can gain access to both individual and collective memories, aiding to a better understanding of our cultural heritage on the one hand, and, on the other hand, restoring missing or forgotten pieces of Europe's oral history.

Keywords: Romanian oral history, Phonogramic Archive, digitalization

Résumé

În această prezentare ne propunem să atragem atenția asupra necesității conservării și digitalizării Arhivei fonogramice a limbii române (AFLR). În momentul de față, AFLR este cea mai bogată și cuprinzătoare colecție de texte dialectale din România. Cu toate acestea, doar o treime din material a fost digitalizat, existând riscul ca benzile rămase să se deterioreze, ducând la pierderea înregistrărilor. Protejarea arhivei contribuie, pe de o parte, la recuperarea narațiunilor individuale și colective, iar, pe de altă parte, la înțelegerea și valorificarea patrimoniului cultural, respectiv redobândirea unor elemente pierdute sau uitate din istoria orală a Europei.

1. Introduction

The aim of this paper is to raise awareness concerning the protection and preservation of the Romanian Phonogramic Archive (AFLR), a repository of the oral history of the country. At the present moment, there is an urgent need to digitalize the Archive in order to prevent further degradation or even possible loss of the entire collection.

2. Romanian – a brief presentation

Romanian is the only surviving Eastern-European Romance language (Rosetti, 1986), being the national language of Romania. It descends from the vernacular variant of Latin, branching into Daco-Romanian, spoken north of the Danube, and three south Danubian dialects, Aromanian, Megleno-Romanian, and Istro-Romanian. The northern dialect is what is usually referred to as Romanian, while the southern tongues have the status of oral dialects, abiding by the literary language of the country they are found in, such as Greek, Bulgarian, Croatian or Albanian. It is generally considered that the Daco-Romanian variety is either divided in two areas, north-west and south-east or it has five subdialects, corresponding to the provinces of Muntenia, Moldova, Crișana, Maramureș and Banat. The north-Danubian subdialects show a remarkable unity, despite external influences, setting Romanian apart from other Romance languages.
Romania joined the European Union in January, 2007, and is registered in the UNESCO World Heritage List with eight properties, six cultural and two natural (https://whc.unesco.org/en/statesparties/ro).

3. The AFLR Archive

AFLR stands out as being the largest collection and most diversified collection of dialectal texts and ethno-linguistic recordings in Romania (over 2000 reel tapes, 30-90 minutes per tape), encompassing material from all Daco-Romanian subdialects, both in the country and across the borders. A monumental work, both in conception, structure and scope, The National Phonogramic Archive is kept at "Iorgu Iordan – Alexandru Rosetti" Institute of Linguistics, Bucharest, Romania (https://www.lingv.ro/).

3.1 Historic overview

The first recordings began in the 1930's, by D. Șandru, in different areas of the country, with modern techniques for that time, representing an important step in Romanian dialectology. The texts transcribed were published between 1933 and 1948. Unfortunately, in 1944, the Bucharest Phonetic Laboratory at the Faculty of Letters was destroyed by air bombings, and along with it the invaluable collection of phonograph recordings of Romanian dialects north of the Danube was lost. After the Second World War, dialectal research in Romania started to flourish. One such key moment in the national development of dialectal research is the initiation (1955-1957) and founding (in 1958) of AFLR, designed as a "sonorous museum", representing all regional varieties of Romanian (Șuteu, 1958).
At the beginning of the 1960's, in parallel with the elaboration of seven regional atlases corresponding to Romania's historical provinces, researchers gathered around the three main university centres, Bucharest, Iași

[1] *Grant de cercetare realizat cu sprijin financiar din Fondul Recurent al Donatorilor, aflat la dispoziția Academiei Române și gestionat prin Fundația „PATRIMONIU” GAR-UM-2019 I-1.5-3; 358c/ 15.10.2019.*

and Cluj-Napoca, wanting to consolidate the phonogramic archives, and field work was resumed. The gathering of dialectal and ethno-texts in order to make up the archives was and continues to be in the centre of attention for dialectologists. The purpose of such collections is to preserve the oral history, traditions, customs and folklore of the population. Dialect investigations not only follow diatopic variation, but also language variation from a sociolinguistic perspective. The AFLR material reflects the speaker's linguistic and social distancing from the mother tongue, the causes and effects of losing national status due to post war treaties reshaping the borders, individual and collective bilingualism, code switching.

Assembled in a rather short time interval (1961-2015), with data collected from a dense network of over 950 villages, the AFLR recordings are representative of Romanian dialects spoken also across the borders (Marin, 1996; Marin, 2003; Marin, 2012). In the 1980's the first recordings took place in America, and in the 1990s, after the fall of the Communist regime, the research intensified, leading to recordings in The Republic of Moldova, certain areas of Ukraine, Bucovina, south-east Hungary, Bulgaria and Serbia. From each investigated community, researchers have systematically interviewed between five and seven speakers from 4 generations, permitting the dialects to be studied in their social dynamics.

3.2 The hidden stories within the Archive

AFLR is an authentic collection of ethno-texts (Neagoe, 1993), vast and original in information regarding traditions, customs and folkloric production. Since Romanian dialect literature is practically non-existent (Vulpe, 1978), these recordings play an important role in linking dialectology with sociolinguistics and the study of folklore.

The Archive includes both descriptive and narrative productions, as well as numerous dialogues between the informants. These oral productions can be classified as: (1) thematic texts – describing customs related to important life events (birth, marriage, death), religious or laic ceremonies, practices of traditional beliefs, specific activities of the informants, (2) spontaneous texts – narrations, historic accounts, memories, anecdotes, jokes, sometimes even songs and carols. At the borderline between the two categories we find folktales, legends, fables and short stories (especially from children). Due to a systematic collection of the data, all of these past narrations are comparable in nature, leading to new lines of inquiry into the oral history of the Romanian language.

3.3 Research topics derived from the Archive

The AFLR recordings contain phonetic productions of paralinguistic information such as changes in intensity, pitch, rhythm, as well as gestural cues, representing an important and original material from a sociolinguistic perspective. AFLR also displays a vast onomasiological inventory that ranges from (nik)names of informants, to names of places accompanied by quite unique explanations delivered by the informants as to the age and origin of the region. It also includes elements of microtoponymy regarding the history of various Romanian rural settlements. Due to its large terminological repertoire (agriculture, manufacturing, viticulture, lumbering, mining, occupations, human anatomy, among others), the material in the Archive is helpful in conducting a structural analysis of dialect terminology.

Important methodological contributions to Romanian dialectology are found in the prefaces of the atlases, regional glossaries, dictionaries and volumes of dialectal texts. Various research topics, theoretical works, and projects have derived from AFLR, such as: (1) collections of dialectal texts written in narrow phonetic transcription gathered from a wider region (TDM, TDO, TDD) or a smaller geographical area (TD–Porţile de Fier, TD–BN, among others); (2) dialectal glossaries focusing not only on the words attested in the transcribed texts, but also itemizing and explaining all regional terms found in the broader recorded area (TDD, TD–BN, TD–Bas., TD–Nistru), (3) dialectal monographs, describing the phonetic, phonological, lexical, morphosyntactic and pragmatic features of different microzones characterized by ethnographic and linguistic unity (TD–Ung., TD–Bulg., TD–Ucraina, TD–Banatul sârbesc); (4) dictionaries – either dedicated to the lexical peculiarities of one village or focused on wider regions such as the *Dictionary of Southern Dacoromanian Varieties* (DGS); (5) anthologies reflecting the richness and the vastness of the dialectal data stored in the Archive, sections of which have been published in works referred to as "sonorous maps", *"Harta sonoră" a graiurilor şi dialectelor limbii române* (Marin and Tiugan, 2014), and "*Harta sonoră" a graiurilor dacoromâne vorbite în afara graniţelor României* (Marin et al., 2016), their novelty residing in the correlation between the dialectal phonetic transcription, the literalization and the digitise tape recordings; (6) studies, articles and PhD theses. Still, only a third of the Archive has been processed so far, with various manuscripts unpublished coupled with the urgent need to digitize the remaining files (either text or audio).

4. Digitising the Archive

Through the preservation of oral archives, we can gain access to both individual and collective memories, aiding to a better understanding of our cultural heritage on the one hand, and, on the other hand, restoring missing or forgotten pieces of Europe's oral history. This being said, we need to take into account both the digitization of the tapes, as well as that of the metadata content (information about gender, age, regional dialect, accompanied by photographs and phonetic transcription of various pronunciation variants belonging to the informant).

4.1 Current progress

The first recordings for AFLR were made on wax cylinders and continued with reel-to-reel tapes, from the 1960's until the 1990's, ensuring a better audio quality. These oral narratives from the past are kept in the Archive either on an audio (reel to reel tapes) or text format (manuscripts, glossaries, dialectal transcripts). At the moment, our internal projects focus mainly on audio digital retrieval

(*Proiectul Consevarea şi gestionarea Arhivei fonogramice a limbii române, Programul IX: BANCA DE INFORMARE ŞI DOCUMENTARE LINGVISTICĂ* [Project Preservation and Management of the Romanian Phonogramic Archive, Program IX: Database for Information and Linguistic Documentation]). Since only 300 recordings have been digitised so far, there is a pending need to develop digital preservation projects so as to prevent further degradation of the entire collection.

4.2 The MIDVAL project

From all recorded villages, dialectologist collected and organised each speaker's metadata file according to social background ("informant's file") and individual pronunciation variants ("phonogramic file"). The resulting hand written documents were then assembled in "village notebooks". Over 950 such "notebooks" exist in the Archive alongside the corresponding reel to reel tapes.

By means of the MIDVAL project ("Metode moderne de instruire și valorificare digitală a documentelor cultural-istorice" [Modern methods for training and digitally restoring cultural-historical documents]), a recent winner of the Romanian Academy funding competition (GAR-UM-2019 I-1.5-3, 2019 – 2021), the metadata files will be organised, properly stored and digitally preserved (https://www.lingv.ro/index.php?option=com_content&view=article&id=354&Itemid=250). This project (coordinated by the main author) aims to implement modern techniques regarding human resource training and digital preservation of the national oral cultural heritage.

5. Hopes for the future

The material found within the National Phonemic Archive, the richness of the ethnographic, linguistic and historic information it discloses, represents cultural assets which need to be valued and protected. In conclusion, our recommendations are the following: (1) the need of a national program in order to create a heritage value-based management and a sustainable framework for future projects, (2) encouraging intercultural dialogue and raising awareness of the importance of past narrations belonging to less represented European languages such as Romanian, (3) digitalizing and to preserving this intangible cultural heritage describing not only the history of an overlooked Romance language, but also gaining a better understanding of Europe's oral history.

6. Bibliographical References

Marin, M. (1996). Arhiva fonogramică a limbii române (După 40 de ani). *Revista de lingvistică şi ştiinţă literară*, 1: 41–46.

Marin, M. (2003). Metode și principii în abordarea graiurilor românești vorbite în medii alogene. M. Berényi (ed.), Simpozion. Comunicările celui de al XI-lea Simpozion al cercetătorilor români din Ungaria (Giula, 23-24 noiembrie 2002), Giula, Research Institute of the Romanians in Hungary, 163–181.

Marin, M. (2012). Arhiva fonogramică a limbii române, sursă de material pentru cercetări interdisciplinare. M. Constantinescu, G. Stoica, O. Uță Bărbulescu (eds.), Modernitate și interdisciplinaritate în cercetarea lingvistică. Omagiu doamnei profesoare Liliana Ionescu-Ruxăndoiu, Bucharest, Bucharest University Press, 342–350.

Marin, M. and Tiugan, M. (2014). *Harta sonoră a graiurilor şi dialectelor limbii române*, Bucharest, Romanian Academy Press (+ CD).

Marin, M., Radu, C. I., Răuţu, D. and Tiugan, M. (2016). *Harta sonoră a graiurilor dacoromâne vorbite în afara graniţelor României*, Bucharest, Romanian Academy Press.

Neagoe, V. (1993). Etnotext – text dialectal. *Fonetică şi dialectologie*, XII, 123–132.

Rosetti. A. (1986). *Istoria limbii române. I. De la origini până la începutul secolului al XVII-lea*, Bucharest, Editura Ştiinţifică si Enciclopedică.

Şuteu, V. (1958). Arhiva fonogramică a limbii române. *Fonetică şi dialectologie*, I, 211–219.

Vulpe, M. (1978). „Romanian Dialectology and Sociolinguistics. *Revue Roumaine de Linguistique, Current Trends in Romanian Linguistics*, A Rosetti and S. Golopenţia Eretescu (eds.), 293-328.

ABBREVIATIONS

DGS – Dicționarul graiurilor dacoromâne sudice, coordinator: Maria Marin, de Ion Ionică, Maria Marin, Anca Marinescu, Iulia Mărgărit, Teofil Teaha, Bucharest, Romanian Academy Press, vol. I, 2009, vol. II, 2010, vol. III, 2011.

TD–Bas. – Maria Marin, Iulia Mărgărit, Victorela Neagoe, Vasile Pavel, Graiuri româneşti din Basarabia, Transnistria, nordul Bucovinei şi nordul Maramureşului. Texte dialectale şi glosar, Bucharest, 2000.

TD–BN – Maria Marin, Marilena Tiugan, Texte dialectale şi glosar. Bistriţa-Năsăud, Bucharest, 1987.

TD–Bulg. – Victorela Neagoe, Iulia Mărgărit, Graiuri dacoromâne din nordul Bulgariei. Studiu lingvistic. Texte dialectale. Glosar, Bucharest, Romanian Academy Press, 2006.

TDD – Paul Lăzărescu, Victorela Neagoe, Ruxandra Pană, Nicolae Saramandu Texte dialectale şi glosar. Dobrogea, Bucharest, 1987.

TDM – Texte dialectale. Muntenia, sub conducerea lui Boris Cazacu, vol. I, de Galina Ghiculete, Paul Lăzărescu, Maria Marin, Bogdan Marinescu, Ruxandra Pană, Magdalena Vulpe, Bucharest, Romanian Academy Press, 1973.

TD–Nistru – Maria Marin, Iulia Mărgărit, Victorela Neagoe, Vasile Pavel, Graiuri româneşti de la est de Nistru. Texte dialectale şi glosar, Bucharest, Romanian Academy Press, 2011.

TDO – Boris Cazacu (coord.), Cornelia Cohuţ, Galina Ghiculete, Maria Mărdărescu, Valeriu Şuteu, Magdalena Vulpe, Texte dialectale. Oltenia, Bucharest, Academy Press, 1967.

TD–Porţile de Fier – Cornelia Cohuț, Magdalena Vulpe, Graiul din zona „Porțile de Fier", I. Texte. Sintaxă, Bucharest, Academy Press, 1973.

TD–Ucraina – Maria Marin, Victorela Neagoe, Iulia Mărgărit, Vasile Pavel, Graiuri româneşti din sud-estul Ucrainei. Studiu lingvistic. Texte dialectale. Glosar, Bucharest, Romanian Academy Press, 2016.

TD–Ung. – Maria Marin, Iulia Mărgărit, Graiuri româneşti din Ungaria. Studiu lingvistic. Texte dialectale. Glosar, Bucharest, Romanian Academy Press, 2005.

Proceedings, pages 358–362
Paris, UNESCO Headquarters, 5-6 December, 2019.

Developing technologies for the documentation and description of the low-resource Uralic languages Zyrian Komi and North Saami

Niko Partanen[1], Thierry Poibeau[2], Michael Rießler[3]
[1]University of Helsinki – niko.partanen@helsinki.fi
[2]CNRS & ENS / PSL & Université Sorbonne nouvelle, Paris – thierry.poibeau@ens.fr
[3]University of Eastern Finland, Joensuu – michael.riessler@uef.fi

Abstract

The Uralic languages are spoken in northern Eurasia, and most of them (with the exception of Finnish, Hungarian and Estonian) are non-national endangered languages with varying official support and resources. Language technology can play a major role in better documenting and describing endangered languages and in making the related workflows faster and more efficient. However, applying contemporary methods effectively in this context remains a challenge. In our own projects, we have developed language technologies focusing on low-resource scenarios, specifically for the two Uralic languages Zyrian Komi and North Saami. In addition to providing an overview of this work, we detail what we see as the remaining obstacles and main challenges for our work. Although we focus on individual languages, our experiments translate to the wider situation of endangered languages in Northern Eurasia and beyond.

Keywords: Zyrian Komi (kpv), North Saami (sme), documentary linguistics, language technology, dependency parsing, OCR

Дженьдӧдӧм текст

Урал кывъяс паськалӧмаӧсь Войвыв Евразияын. На пиысь унджыкыс (финн, эст да венгр кындзи) ӧнія кадӧ вошан выйынӧсь. Кыв технологияяс вермасны документируйтны кывъяссӧ, а сідзкӧ отсаласны видзны найӧс дзикӧдз вошӧмысь. Дерт, кокнимӧдасны и таköд йитӧдын вӧчан муköд уджсӧ. Но колӧ пасйыны, мый ӧнія кадся методъясӧн вӧдитчӧны сьӧкыда на. Асланым проектын ми лӧсьӧдім коми-зыран да войвыв саами урал кывъяслы технологияяс. Ӧтувъя серпас петкӧдлӧмысь кындзи тані ми видлалам и сійӧ мытшӧдъяссӧ, кодъяскӧд паныдасим удж нуӧдігӧн. Кӧть ми сёрнитам торъя кывъяс йылысь, миян опыт петкӧдлӧ Войвыв Евразияса уна кывлысь серпассӧ.

1. Introduction

The Uralic languages form a family of languages spoken by approximately 25 million people, predominantly in north-eastern Europe and western Siberia. With the exception of Finnish, Hungarian and Estonian, all Uralic languages are endangered. In this context, language technologies can play a major role in better documenting and describing these languages. Developing language resources and technologies ensures that knowledge about their specificities will be retained and thereby enables us to help in preserving and teaching them, with information technology bringing major benefits to this end.

Originally, language technologies for Uralic languages have been developed only for written language variants (especially by the research groups Giellatekno and Divvun at the University of Tromsø). Our approach – informed by both computational linguistics and (fieldwork-based) documentary linguistics – also includes spoken language data. We have been developing technologies for Komi and Saami based on the most recent advances in natural language processing, and have applied them to a context where annotated corpus data and other resources are lacking.

In this paper, we give an overview of our most recent research (see more specifically Lim et al. (2018b); Partanen et al. (2018b) and Partanen et al. (2018a)) and we detail what are, in our view, the main challenges for language technologies in low-resource language documentation contexts in general. Our publications are focused on two specific languages but we think that a large variety of small languages can be approached in a similar way, especially when raw digitized texts are already available but other kinds of resources, specifically annotated data, are lacking. In fact, this is true for most Uralic languages as well as for several Northern Eurasian languages from other families.

As the lack of annotated data is often emphasized as an obstacle, specific attention must be paid in this context to solutions that enable the rapid increase of annotated data. For most Uralic languages, the problem is not the lack of data, as material has been collected and published in most of these languages since at least the beginning of the 20th century (some of the textual materials are even old enough to be in the Public Domain, making it legally possible to create entirely open and easy reusable datasets). The problem is the lack of annotation for this data.

It must be noted that Zyrian Komi and North Saami are relatively well described and linguistically understood languages with speaker numbers ranging from about 30,000 (North Saami) to 160,000 (Zyrian Komi). Both languages also receive official support, have well-established written norms and are regularly used in media and education, even in vocational and higher academic education. Retrieving textual data for corpus building and (written) language technology is also comparably easy. In fact, extensive written corpora have been created for both languages already: the SIKOR North Saami free corpus with over 30M tokens by Divvun/Giellatekno and the Корпус Коми языка (“Corpus of the Komi language”) with over 50M tokens by the Centre for Innovative Language Technology in Syktyvkar, Komi Republic. Whereas these two corpora are tagged using rule-based NLP, various new approaches in language technology

are also evolving. A recent evaluation of currently available language technology for Finnish (Pirinen, 2019), for instance, showed that the Turku Neural Parser Pipeline clearly outperforms the older rule-based systems for Finnish, and when enough annotated data becomes available, the same can also be expected for other Uralic languages.

2. Summary of the research done so far

We have mainly developed our research along three different directions. One is *the integration of various language technologies in order to get more efficient NLP workflows for fieldwork-based language documentation* (Gerstenberger et al., 2017a; Gerstenberger et al., 2017b). The achievements in corpus building sketched in the section above concern only written language; fieldwork data representing spoken language has not been included in relevant projects earlier. A central approach for us has been to find ways to use language technology so that language documenters – collecting new data in the field or working with legacy data in archives – can work faster and more reliably with their language data. Ideally this should be done in connection with the language documentation activities that would normally take place anyway, and using the tools the researchers are already familiar with. This is also related to the ability to preserve and reuse the same material later on, although there are needs for improvement in all levels, from data management and archiving to final publication and re-use in research.

Compared to traditional NLP workflows for written texts, ours must integrate speech technologies, for speech transcription or signal analysis. As for written texts, processing workflows may go beyond pure NLP, so as to integrate document analysis and OCR when it comes to corpus integration of earlier documents (Blokland et al., 2019).

We have also conducted several case studies about *dependency parsing in these low-resource scenarios* (Lim et al., 2018b; Partanen et al., 2018b). Dependency parsing is now a relatively mature technology, mainly based on advanced machine learning techniques that require large amounts of annotated data to get accurate results. This is a major issue for low resource scenarios, but recent techniques based on multilingual models and language transfer have made it to possible to get working results even in extremely low resource scenarios. With Uralic languages the most obvious approaches for multilingual systems would consider closely related languages and contemporary contact languages, and our experiments have covered both. The results are of course far from perfect but our aim in the long run is of course to use these methods for language documentation. Automatic annotations need to be revised and corrected, but they are useful to kick-start the annotation process and they also make it possible to considerably increase the size of the data produced (which, in turn, makes it possible to train better parsers that will require less manual correction).

The third portion of our work has focused more into *concrete resource creation, which is illustrated by two Zyrian Komi treebanks* (Partanen et al., 2018b) and a large spoken language corpus (Blokland et al., 2020). This shows that technical advances work hand in hand with the production of resources and help maintain and document endangered languages. Our work aligns closely with observations others have made in relation to this field, namely that even a small amount of annotated data still brings at the moment clear improvements into any multilingual scenarios (Meechan-Maddon and Nivre, 2019). As our own datasets have grown, we are replicating and extending our earlier experiments, with the goal of reaching a workable solution for our continuous language documentation work.

Although we have been successful in integrating language technology into language documentation workflows (Gerstenberger et al., 2017a), there are still numerous open questions about how the whole infrastructure should be set up so that resources and applications would be most beneficial for both field linguists and computational linguists. Some of these open questions are discussed next.

2.1. Persistent archiving of language documentation corpora

In the last 20 years a large number of language documentation projects have been conducted all over the world and provided vast digital resources on endangered and previously undocumented languages. There are, however, numerous problems in the actual use of these materials, especially in more computerized workflows. Language documentation projects produce complex multimedia collections and associated metadata. Lots of attention has been paid to open and shared formats in language documentation (Seyfeddinipur et al., 2019). Still, it is often a major challenge to maintain long-term consistency in such collections. As an outcome, language documentation corpora may be unsystematic in ways that make reusing them difficult. This relates closely to the fact that work on individual languages often continues for years, even decades: for this reason, work practices within a language documentation project have to be thoroughly documented, so that even changing and entirely new teams can connect and continue previous work.

A solution in our own projects is to define the intended data structures in a machine readable format, and to build a set of tests that continuously validate that both structure and content are within expected definitions (Partanen, 2019a).

2.2. Text recognition

A large number of linguistic resources for endangered languages, representing transcribed spoken language, have been published in books or are stored as manuscripts. For many publications audio recordings underlying the transcripts are stored in various private or public archives. The usefulness of these data for future work with endangered languages is without question (Blokland et al., 2019). However, merging this analogue data into digitally-born corpora can be challenging. Finding the relevant audio files and digitizing them as well as converting rare and non-standard writing systems designed (e.g. variants of various phonetic alphabets typically used in the printed texts) into contemporary standard orthography can be such challenges. Much more problematic, however, is often the exact matching between the transcript and the original audio. A reason for this seems to be that the recorded speech may be fuzzy and subject to interpretation and the later published version has

gone through orthographic and stylistic editing without taking the original recording into account.

So far our own work with such legacy resources has focused on digitizing the relevant texts, building OCR models for rendering the different original scripts and integrating the resulting data into our corpus infrastructure. The alignment of the processed texts with the original audio, if available, is an upcoming task.

The tools for performing text recognition in itself have improved considerably during the last years. Only a few years ago alternatives to ABBYY FineReader were relatively few, even though the problems present with this commercial software are numerous (Partanen, 2017). Recently it has been possible to train very well performing OCR models with open source software, such as Tesseract, Ocropy and Calamari (Partanen and Rießler, 2019). In connection to this kind of work we have published Ground Truth datasets and OCR models (Partanen and Rießler, 2019; Partanen, 2019b). The best practices in sharing Ground Truth data need to be taken into account, for example, by following the conventions used by the National Library of Finland (Kettunen et al., 2018). However, after the texts have been retrieved from the documents, several issues remain in their successful transliteration and normalization. There has been very promising work on normalizing Finnish dialect texts as a character level machine translation task, with achieved word error rate in around 5% (Partanen et al., 2019). More work is acutely needed in whether such approaches are viable also with endangered languages and when less data is available. Same normalization problem, however, also in language documentation context.

2.3. Dependency parsing

Since 2017 a number of experiments have been carried out by our team with dependency parsing of low-resource languages. The system used in these experiments is the Multilingual BIST parser (Lim and Poibeau, 2017). The experiments were done with cross-lingual scenarios where data from related languages and contact languages were used alongside the minimal training data in the target language (Lim et al., 2018b; Lim et al., 2018a). The experiments were promising, and the LAS score on Northern Saami was 51.54 and for Zyrian Komi 56.66. This improved from the parsing results demonstrated for Northern Saami in CoNLL 2017 Shared Task (Lim et al., 2018b, p. 2233). Despite improvements the achieved performance was generally not high enough for practical applications. Additional experiments were done with code-switching data, in order to understand how well this kind of a multilingual system is able to parse data that contains both of the languages it was trained on (Partanen et al., 2018b). It has to be noted that in spoken data, such as typically included in spoken corpora of endangered languages, code-switching is the rule rather than the exception. Any work with dependency parsing of such data needs to consider this in order to be applicable in real-world tasks. Our results showed no major differences between the monolingual and multilingual test sets, which, however, remains open to further analysis as the test data was very small.

In order to create the foundation for more comprehensive testing of different NLP methods, two Zyrian Komi treebanks (Partanen et al., 2018a) were created as part of the Universal Dependencies (UD) project (Nivre et al., 2018). We believe these treebanks will also become important resources for linguistic research, beyond computational linguistics. Increasing their size and coverage has been a continuous effort since the beginning and resulted in various updates and releases. Part of this work has also focused on comparing the UD treebanks across different other Uralic languages, in order to ensure that the cross-linguistic treebank data remain comparable in the future, especially when new treebanks are added and the annotation scheme is taken into use in a new language (Partanen and Rueter, 2019).

Recent work of Lim et al. (2020 accepted) presents good results on dependency parsing with the use of semisupervised learning. A small initial training treebank is appended with a larger amount of plain text, which is used to learn a meta structure that improved LAS scores even by 9.3 points. If such improvement could be seen also in scenarios we have tested in our previous papers, we would be approaching the point where the result could be useful for documentary linguistics working with fieldwork data. This is particularly interesting since mixing small manually tagged data sets with larger amounts of untagged text fits exactly the scenario described in this paper: both Komi and Saami have very large non-annotated corpora and relatively small manually created resources.

Since the latest UD release contains Karelian (closely related to Finnish), Skolt Saami (closely related to North Saami) and Permiak Komi (closely related to Zyrian Komi) treebanks, the possibilities of multilingual dependency parsing between new very closely related languages is increasingly becoming possible to investigate.

3. Conclusions

This paper presented our ongoing work using language technology for better linguistic documentation and description of Zyrian Komi and North Saami. However, language technology can also be applied in practical projects aiming at language revitalization and language maintenance. Therefore, building any kind of language technology for an endangered language is potentially of relevance for speakers and learners of endangered languages as well as for language planners. Building language technology while paying attention to Open Source technologies and datasets ensures that at least the results will be available for the community and potentially reusable in practical applications in the future.

From the perspective of reusability, persistent archiving is very central to computational workflows in documentary linguistics, especially if multimedia data is included. Unfortunately, best practices around regularly and automatically updating the archived collections have yet to be established. Persistent identifiers are used, but conventions such as semantic versioning are still rare. Archived materials usually cannot be updated through an API, which makes it difficult to interact with the collections and to link their maintenance to more automatized workflows, like the ones we use in our projects.

Acknowledgements

The research summarized in this paper has been funded by various organisations, among them the German Science Foundation, Kone Foundation, Paris Sciences et Lettres, RGNF-CNRS, and Volkswagen Foundation. Thierry Poibeau is supported by a Prairie (Paris Artificial Intelligence Research Institute) fellowship. Thanks to Vasily Chuprov for translating the abstract into Komi.

4. Bibliographical References

Blokland, R., Partanen, N., Rießler, M., and Wilbur, J. (2019). Using computational approaches to integrate endangered language legacy data into documentation corpora: Past experiences and challenges ahead. In *Workshop on Computational Methods for Endangered Languages, Honolulu, Hawai'i, USA, February 26–27, 2019*, volume 2, pages 24–30. University of Colorado.

Blokland, R., Fedina, M., Partanen, N., and Rießler, M. (2020). Spoken Komi Corpus. The Language Bank of Finland version.

Gerstenberger, C., Partanen, N., and Rießler, M. (2017a). Instant annotations in ELAN corpora of spoken and written komi, an endangered language of the Barents Sea region. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 57–66.

Gerstenberger, C., Partanen, N., Rießler, M., and Wilbur, J. (2017b). Utilizing language technology in the documentation of endangered Uralic languages. *Northern European Journal of Language Technology: Special Issue on Uralic Language Technology*.

Kettunen, K., Kervinen, J., and Koistinen, M. (2018). Creating and using Ground Truth OCR sample data for Finnish historical newspapers and journals. In *Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference*, pages 162–169.

Lim, K. and Poibeau, T. (2017). A system for multilingual dependency parsing based on bidirectional lstm feature representations. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 63–70.

Lim, K., Partanen, N., and Poibeau, T. (2018a). Analyse syntaxique de langues faiblement dotées à partir de plongements de mots multilingues. *Traitement Automatique des Langues*, 59(3):67–91.

Lim, K., Partanen, N., and Poibeau, T. (2018b). Multilingual dependency parsing for low-resource languages: Case studies on North Saami and Komi-Zyrian. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Lim, K., Yoon Lee, J., Carbonell, J., and Poibeau, T. (2020 accepted). Semi-supervised learning on meta structure: Multi-task tagging and parsing in low-resource scenarios. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Meechan-Maddon, A. and Nivre, J. (2019). How to parse low-resource languages: Cross-lingual parsing, target language annotation, or both? In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 112–120.

Nivre, J., Abrams, M., Agić, Ž., Ahrenberg, L., Antonsen, L., Aranzabe, M. J., Arutie, G., Asahara, M., Ateyah, L., Attia, M., Atutxa, A., Augustinus, L., Badmaeva, E., Ballesteros, M., Banerjee, E., Bank, S., Barbu Mititelu, V., Bauer, J., Bellato, S., Bengoetxea, K., Bhat, R. A., Biagetti, E., Bick, E., Blokland, R., Bobicev, V., Börstell, C., Bosco, C., Bouma, G., Bowman, S., Boyd, A., Burchardt, A., Candito, M., Caron, B., Caron, G., Cebiroğlu Eryiğit, G., Celano, G. G. A., Cetin, S., Chalub, F., Choi, J., Cho, Y., Chun, J., Cinková, S., Collomb, A., Çöltekin, Ç., Connor, M., Courtin, M., Davidson, E., de Marneffe, M.-C., de Paiva, V., Diaz de Ilarraza, A., Dickerson, C., Dirix, P., Dobrovoljc, K., Dozat, T., Droganova, K., Dwivedi, P., Eli, M., Elkahky, A., Ephrem, B., Erjavec, T., Etienne, A., Farkas, R., Fernandez Alcalde, H., Foster, J., Freitas, C., Gajdošová, K., Galbraith, D., Garcia, M., Gärdenfors, M., Gerdes, K., Ginter, F., Goenaga, I., Gojenola, K., Gökırmak, M., Goldberg, Y., Gómez Guinovart, X., Gonzáles Saavedra, B., Grioni, M., Grūzītis, N., Guillaume, B., Guillot-Barbance, C., Habash, N., Hajič, J., Hajič jr., J., Hà Mỹ, L., Han, N.-R., Harris, K., Haug, D., Hladká, B., Hlaváčová, J., Hociung, F., Hohle, P., Hwang, J., Ion, R., Irimia, E., Jelínek, T., Johannsen, A., Jørgensen, F., Kaşıkara, H., Kahane, S., Kanayama, H., Kanerva, J., Kayadelen, T., Kettnerová, V., Kirchner, J., Kotsyba, N., Krek, S., Kwak, S., Laippala, V., Lambertino, L., Lando, T., Larasati, S. D., Lavrentiev, A., Lee, J., Lê Hồng, P., Lenci, A., Lertpradit, S., Leung, H., Li, C. Y., Li, J., Li, K., Lim, K., Ljubešić, N., Loginova, O., Lyashevskaya, O., Lynn, T., Macketanz, V., Makazhanov, A., Mandl, M., Manning, C., Manurung, R., Mărănduc, C., Mareček, D., Marheinecke, K., Martínez Alonso, H., Martins, A., Mašek, J., Matsumoto, Y., McDonald, R., Mendonça, G., Miekka, N., Missilä, A., Mititelu, C., Miyao, Y., Montemagni, S., More, A., Moreno Romero, L., Mori, S., Mortensen, B., Moskalevskyi, B., Muischnek, K., Murawaki, Y., Müürisep, K., Nainwani, P., Navarro Horñiacek, J. I., Nedoluzhko, A., Nešpore-Bērzkalne, G., Nguyễn Thị, L., Nguyễn Thị Minh, H., Nikolaev, V., Nitisaroj, R., Nurmi, H., Ojala, S., Olúòkun, A., Omura, M., Osenova, P., Östling, R., Øvrelid, L., Partanen, N., Pascual, E., Passarotti, M., Patejuk, A., Peng, S., Perez, C.-A., Perrier, G., Petrov, S., Piitulainen, J., Pitler, E., Plank, B., Poibeau, T., Popel, M., Pretkalniņa, L., Prévost, S., Prokopidis, P., Przepiórkowski, A., Puolakainen, T., Pyysalo, S., Rääbis, A., Rademaker, A., Ramasamy, L., Rama, T., Ramisch, C., Ravishankar, V., Real, L., Reddy, S., Rehm, G., Rießler, M., Rinaldi, L., Rituma, L., Rocha, L., Romanenko, M., Rosa, R., Rovati, D., Ro□ca, V., Rudina, O., Sadde, S., Saleh, S., Samardžić, T., Samson, S., Sanguinetti, M., Saulīte, B., Sawanakunanon, Y., Schneider, N., Schuster, S., Seddah, D., Seeker, W., Seraji, M., Shen, M., Shimada, A., Shohibussirri, M., Sichinava, D., Silveira, N., Simi, M., Simionescu, R., Simkó, K., Šimková, M., Simov, K., Smith, A., Soares-Bastos, I., Stella, A., Straka, M., Str-

nadová, J., Suhr, A., Sulubacak, U., Szántó, Z., Taji, D., Takahashi, Y., Tanaka, T., Tellier, I., Trosterud, T., Trukhina, A., Tsarfaty, R., Tyers, F., Uematsu, S., Urešová, Z., Uria, L., Uszkoreit, H., Vajjala, S., van Niekerk, D., van Noord, G., Varga, V., Vincze, V., Wallin, L., Washington, J. N., Williams, S., Wirén, M., Woldemariam, T., Wong, T.-s., Yan, C., Yavrumyan, M. M., Yu, Z., Žabokrtský, Z., Zeldes, A., Zeman, D., Zhang, M., and Zhu, H. (2018). Universal dependencies 2.2. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Partanen, N. and Rießler, M. (2019). An OCR system for the Unified Northern Alphabet. In *Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages*, pages 77–89.

Partanen, N. and Rießler, M. (2019). langdoc/unified-northern-alphabet-ocr: Unified Northern Alphabet OCR Ground Truth, March.

Partanen, N. and Rueter, J. (2019). Survey of Uralic universal dependencies development. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 78–86.

Partanen, N., Blokland, R., Lim, K., Poibeau, T., and Rießler, M. (2018a). The first Komi-Zyrian Universal Dependencies treebanks. In *Second Workshop on Universal Dependencies (UDW 2018), November 2018, Brussels, Belgium*, pages 126–132.

Partanen, N., Lim, K., Rießler, M., and Poibeau, T. (2018b). Dependency parsing of code-switching data with cross-lingual feature representations. In *Proceedings of the Fourth International Workshop on Computational Linguistics of Uralic Languages*, pages 1–17, jan.

Partanen, N., Hämäläinen, M., and Alnajjar, K. (2019). Dialect text normalization to normative standard Finnish. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 141–146.

Partanen, N. (2017). Challenges in OCR today: Report on experiences from INEL. In *Elektronnaja pismennost narodov Rossijskoj Federacii: opyt, problemy i perspektivy*, pages 263–273.

Partanen, N. (2019a). langdoc/elan-tests: Language documentation corpus validation scripts, December.

Partanen, N. (2019b). nikopartanen/vyl-tujod-ocr: Vyl' Tujöd newspaper Ground Truth, May.

Pirinen, T. A. (2019). Neural and rule-based Finnish NLP models—expectations, experiments and experiences. In *Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages*, pages 104–114.

Seyfeddinipur, M., Ameka, F., Bolton, L., Blumtritt, J., Carpenter, B., Cruz, H., Drude, S., Epps, P. L., Ferreira, V., Galucio, A. V., Hellwig, B., Hinte, O., Holton, G., Jung, D., Buddeberg, I. K., Krifka, M., Kung, S., Monroig, M., Neba, A. N., Nordhoff, S., Pakendorf, B., von Prince, K., Rau, F., Rice, K., Rießler, M., Brenig, V. S., Thieberger, N., Trilsbeek, P., van der Voort, H., and Woodbury, T. (2019). Public access to research data in language documentation. *Language Documentation & Conservation*, 13:545–563, oct.

Proceedings, pages 363–366
Paris, UNESCO Headquarters, 5-6 December, 2019.

Development of Technology for Indian Languages: Indian Government Initiatives

Sunil Kumar Srivastava

Government of India
Ministry of Electronics and Information Technology
Electronics Niketan, 6, Lodi Road, New Delhi - 110003
sks@meity.gov.in

Abstract

With twenty two constitutionally recognized languages written in twelve different scripts and several hundred dialects, India faces a major challenge when it comes to the issue of language. Under Technology Development for Indian Languages (TDIL) Programme, Government of India has sponsored several projects which has led to development of technology and tools in the areas like machine translation, automatic speech recognition, optical character recognition, etc. Government is now initiating Natural Language Translation Mission which aims at building and deploying operational machine translation systems for major Indian languages.

बारह अलग-अलग लिपियों में लिखी जाने वाली और कई सौ बोलियों में बोली जाने वाली बाईस संविधान द्वारा मान्यता प्राप्त भाषाओं के प्रयोग में होने के कारण, भारत को भाषा के बिंदु पर एक बड़ी चुनौती का सामना करना पड़ता है। भारतीय भाषाओं के लिए प्रौद्योगिकी विकास (टीडीआईएल) कार्यक्रम के तहत, भारत सरकार ने कई परियोजनाओं को प्रायोजित किया है, जिसके कारण मशीन आधारित अनुवाद, आटोमैटिक स्पीच रिकग्निशन, आप्टिकल कैरेक्टर रिकग्निशन, आदि क्षेत्रों में प्रौद्योगिकी और उपकरणों का विकास हुआ है। अब भारत सरकार प्राकृतिक भाषाओं के अनुवाद पर एक मिशन की शुरुआत कर रही है। इस मिशन का उद्देश्य प्रमुख भारतीय भाषाओं के लिए मशीन आधारित अनुवाद प्रणालियों का विकास करना और प्रयोग में लाना है।

Keywords: language technology, machine translation, India

1. Introduction

India is a country of diversity in several aspects including languages. With 22 constitutionally recognized languages written in 12 different scripts and several hundred dialects, the country faces a challenge when it comes to the issue of communication. During the recent past, India has seen an emergence of digital economy. A number of government services are being offered in digital form - ranging from birth certificate to filing petitions in the courts. Almost all kinds of organizations, public or private, commercial or non-commercial have made their presence felt in the digital space. However, one major challenge is the access to information and services in the native languages. About 15% people only can speak and write in English. The remaining population is unable to derive the benefits of IT as most of the solutions developed have interfaces in English. Language becomes a barrier for a large percentage of population.

Government of India has been taking steps towards the development of technology for Indian languages since the eighties. In early nineties, it initiated an R&D Programme titled *Technology Development for Indian Languages* [4]. During the beginning years, the programme was primarily concerned with the development of standards for Indian languages, device drivers for Indian languages, fonts for various scripts, and localization of the popular open-source tools such as Linux, OpenOffice, etc. These tools were distributed across the country. It was especially useful for those languages where tools were not available from the vendors.

Later on, the focus shifted to technology development in several new areas including, but not limited to, automatic speech recognition (ASR), text-to-speech synthesis (TTS), optical character recognition (OCR) and machine translation (MT). Several projects were sponsored to the academic/R&D institutions in consortium mode where one institution was leading the R&D work in that area (e.g. ASR) and other member institutions were working on the development of technology for specific languages. Some of the lead institutions in the respective areas were IIT Madras (ASR and TTS), IISc Bangalore (OCR), IIIT Hyderabad (MT from Indian languages to Indian languages) and C-DAC Pune (MT from English to Indian languages). The achievements in the individual areas are briefly described below.

2. Status

Several demonstration level prototypes have been developed in the above-mentioned areas. Some of the tools such as TTS have been used by the developers in applications. These are briefly discussed below.

2.1. Automatic Speech Recognition (ASR)

Under the TDIL Programme, IIT Madras has been working on the development of large vocabulary continuous speech recognition (LVCSR) systems. It has developed a system called, Mandi which provides speech-based access to agricultural commodity prices and weather information in 11 Indian languages/dialects. This system provides current commodity prices in local markets and local weather information to users in a convenient manner. It takes the data available on AgmarkNet and AgriMet websites. The system has been built from scratch using open-source tools/software so that it can be used by by public institutions with no licensing issue or cost.

2.2 Text-to-Speech (TTS) Synthesis

Text-To-Speech (TTS) synthesis system for 10 Indian languages viz. Tamil, Telugu, Marathi, Bodo, Kannada, Odia, Hindi, Malayalam, Manipuri & Rajasthani have been developed under a consortium project under the leadership of IIT Madras. Mean Opinion Score (MOS) of these TTS systems on the scale of 0 to 5 is 3.2 or more. It has also been made available in open source under creative commons (CC-BY 4.0) license and can be downloaded from [1]. One of the developed systems called m-Vachak, helps visually challenged people in accessing digital information. The system has also been made available on Android based operating system, Indus OS. As of now, there are 1.54 million activations on 8 Mobile Brands (Micromax, Celkon, Swipe, Karbonn, Intex, Trio, Sansui & Datawind) supporting Indus OS. Applications like browser plugin for Mozilla Firefox and Google Chrome, SMS Reader for Indian languages, TTS voices have been made available to the public through TDIL Data centre [3] for use and feedback.

2.3 Machine Translation (MAT)

2.3.1 Indian Language to Indian Language Machine Translation System: The system called *Sampark*, uses both rules-based and dictionary-based algorithms with statistical machine learning approach. It has been developed for 18 language pairs[Hindi↔Punjabi, Hindi↔Urdu, Hindi↔ Tamil, Hindi↔ Tamil , Hindi↔ Bengali, Hindi↔ Marathi, Hindi↔Kannada, Telugu↔Tamil, Malayalam↔Tamil]. The system was developed by a consortium of institutions under the leadership of IIIT Hyderabad.

2.3.2 English to Indian Languages Machine Translation System (AnglaMT): The system called AnglaMT, is a rule based machine translation system for English to 8 Indian languages [English↔ Hindi, English↔ Malayalam, English↔ Bengali, English↔ Urdu, English↔ Punjabi, English↔ Tamil, English↔ Assamese, English↔ Nepali] developed by a connsortum of institutions under the leadership of CDAC Noida.

2.3.3 English to Indian Languages Machine Translation System (Anuvadaksh): The system called Anuvadaksh uses statistical and example-based machine translation techniques for translation from English to 8 Indian languages [English↔ Hindi, English↔ Marathi, English↔Bengali, English↔Urdu, English↔Odia, English↔ Tamil, English↔Gujarati, English↔Bodo]. It was developed by a consortium of institutions under the leadership of CDAC Pune.

2.4. Optical Character Recognition (OCR)

OCR system has been developed for 13 Indian languages- Assamese, Bangla, Gurmukhi, Hindi, Kannada, Malayalam, Tamil, Telugu, Urdu, Gujarati, Oriya, Manipuri and Marathi. The preprocessing routines such as adaptive binarization, noise cleaning, skew corrections routines were developed by different consortium partners. Different classifiers such as SVM, KNN, LSTM were used. These can be used under Windows, Linux and Web version.

2.5. Language Technology Tools

A large number of tools have been developed for all major Indian languages. These tools have been used in several applications for Indian languages. The centres have also developed linguistic resources like dictionaries, taggers, spell checkers, CLDR, grammar checkers, sorting utilities, thesauri, tagged lexicons, and information extraction & retrieval and standards.

2.6. National Platform for Language Technology

The outcomes of the projects undertaken under TDIL Programme have been showcased at the Indian Language Technology Proliferation & Deployment Centre portal. This portal has been acting as a national repository for linguistic resources, tools and applications being developed under the various TDIL sponsored projects. Now, the portal

has been redesigned and is being launched as National Platform for Language Technology (http:nplt.gov.in). The portal will work as an e-marketplace for linguistic resources and tools.

3. Natural Language Translation Mission

During March 2019, nine national science & technology missions were announced by the Principal Scientific Adviser to Government of india. These missions have been recommended by Prime Minister's Science, Technology & Innovation Advisory Council (PM-STIAC) [2].

3.1. Objectives:

The objectives of the mission are the following:

3.1.1. To build a high-quality speech to speech machine translation (SSMT) system for major Indian languages;

3.1.2. To create and nurture an ecosystem involving start-ups, central/state government agencies working together to develop and deploy innovative products and services in Indian languages;

3.1.3. To increase the content in Indian languages on Internet substantially in the domains of public interest, particularly science & technology, education, healthcare, governance, and law & justice, etc.

3.2. Implementation Strategy:

There are five elements in the mission: R&D consortia, CoEs, start-ups, sub-missions and National Hub for Language Technology (NHLT). Under TDIL Programme, several R&D consortia were created to develop technology for Indian languages and these have been working for almost a decade. In the present mission, four consortia will be working on the respective areas viz. Speech Technology(ASR & TTS), English to Indian Languages Machine Translation (EILMT), Indian language to Indian language Machine Translation (ILMT), and Optical Character recognition (OCR). The consortia will work towards upgrading the technology and will also provide technical assistance to CoEs, start-ups, etc.

CoEs, the second element will be responsible for translating the lab prototypes into commercial products. It has been seen that the academic institutions have not been able to take the lab prototypes from the lab to the land as their focus has been on R&D. These centers, essentially, would be Center of Engineering which will do necessary software engineering for converting the lab prototype into a commercial product. CoEs will also provide support to the developers, especially start-ups who would use the technology to develop solutions to meet the requirements of the users.

Start-ups are the third strategic element in the mission. The start-ups are being envisaged to be primary vehicle for developing the applications and providing services in language technology space. These will also be used to create the required high volumes of language data. The academic institutions will provide technical guidance to the start-ups in the process of resource creation.

Sub-Missions on individual languages are the fourth element of the Mission. These will be launched with the participation of the states. Each sub-mission will focus on one of the recognized languages. The sub-mission will be handled by the state where it is used. In case of the languages which are spoken across several states, all the states will be participating. In order to increase the content in Indian languages on the internet, the content available on the Internet will be translated into Indian languages using machine translation systems followed by review by human translators. Once corrected, the content will be made available for training of the machine translation systems.

Finally, it is proposed to create a National Hub for Language Technology (NHLT) to provide services and central facilities including National Machine Translation Service through Bahu-Bhashak Platform, National Language Technology Platform for resource sharing. NHLT will also be responsible for conducting Grand-Challenges and contests in the area of language technology.

4. Conclusions

The article has described the activities which have been undertaken towards the development of technology for Indian languages. Efforts have been made since the nineties. However, though some good prototype systems have been developed, very few have reached to the end users. In order to make use of the advances in language technology space, Government of India is initiating a mission on natural language translation which aims at

developing machine translation systems for all major Indian languages.

5. Bibliographical References

1. Indic TTS Project. https://www.iitm.ac.in/donlab/tts/

2. PMSTIAC - Missions. http://psa.gov.in/pmstiac-missions

3. TDIL Data Centre. http://www.tdil-dc.in

4. TDIL Programme. [http://tdil.meity.gov.in/

Proceedings, pages 367–370
Paris, UNESCO Headquarters, 5-6 December, 2019.

Planning for Language Technology Development and Language Revitalization in Wales

Delyth Prys, Dewi Bryn Jones, Gruffudd Prys
Language Technologies Unit, Bangor University, Wales
{d.prys, d.b.jones, g.prys}@bangor.ac.uk

Abstract

Welsh speakers have striven to maintain and revitalize their language in the face of the increasing dominance of English. Language technologies, the internet and digital media important to make Welsh more attractive, relevant and accessible. From the initial efforts of a few academics, key resources and tools were created and formed into coherent building blocks to be reused and refined, keeping costs low and working with different stakeholders, including government, industry and enthusiasts. Recent publication of a government LT Action Plan has enabled longer term planning in ways which might also interest other language communities.

Keywords: language technology planning, best practice, Welsh

Résumé

Mae siaradwyr Cymraeg wedi ceisio cynnal ac adfer eu hiaith yn wyneb bygythiad cynyddol y Saesneg. Mae technolegau iaith, y rhyngrwyd a chyfryngau digidol yn bwysig i wneud y Gymraeg yn fwy deniadol, perthnasol a hygyrch. O ymdrechion cychwynnol rhai academyddion, crëwyd adnoddau ac offer allweddol a'u llunio yn flociau adeiladu i'w hailddefnyddio a'u gwella, gan gadw costau yn isel a gweithio gyda gwahanol randdeiliaid, gan gynnwys llywodraeth, diwydiant a charedigion yr iaith. Mae cyhoeddi Cynllun Gweithredu TI y llywodraeth yn ddiweddar wedi galluogi cynllunio tymor hirach mewn ffyrdd a allai fod hefyd o ddiddordeb i gymunedau ieithoedd eraill.

1. Introduction

Under-resourced languages are, by definition, lacking in adequate resources to fulfil their technological needs. It is common in such situations to start working on whatever the most urgent problem happens to be, such as a word list, bilingual dictionary, spell-checker or some other tool or resource. Before long, a handful of different tools and resources might have been created. They can fulfil a short-term need, but in order to be useful in the longer term, it helps that issues such as appropriate licences, sustainable repositories and modular design are all thought through as early as possible.

In the case of Welsh, there was an awareness of some of these issues from an early date, and attempts were made to future-proof the work in low-cost and sustainable ways.

Welsh is spoken by approximately 562,000 people (Office of National Statistics, 2012), and is variously described as minoritized or endangered, or more recently, in digital environments, a less- or under-resource language. After more than a century-long period of decline, there are concerted efforts, by Welsh-speakers, their non-Welsh speaking compatriots, and government, to reverse the decline and ensure the future of Welsh as a spoken, vibrant language, fit for life in the twenty first century.

2. Strategic documents

Since the devolution of power to the National Assembly for Wales in 1999, the government of Wales has published a succession of strategic documents to revitalize Welsh. These include Iaith Pawb: A National Action Plan for a Bilingual Wales (2003); A living language: a language for living (2012) and Cymraeg 2050: Welsh Language Strategy (2017). The latest of these has the ambitious target of nearly doubling the number of Welsh speakers to one million by the year 2050.

All these documents include sections on language technologies and their importance for the revitalization of Welsh. These sections have become increasingly focused and detailed as the technologies themselves have developed and assumed an increasingly central role in our lives over the last twenty years. Cymraeg 2050 emphasizes regional economic development and investing in entrepreneurship programmes to support Welsh speakers in rural areas. Digital technologies have a whole section devoted to them in this document and they also feature prominently in the section on linguistic infrastructure.

The Welsh language and technology have also been integrated into other strategic documents, policies and legislation, so that they do not exist in an isolated environment. Foremost amongst these has been the Well-being of Future Generations (Wales) Act (2015) with its vision to improve the social, economic, environmental and cultural well-being of the people of Wales, and the North Wales Growth Deal (2019) which names Technology as one of its priorities.

A Welsh Language Technology Action Plan (Welsh Government, 2018) added further detail these strategies, outlining "How we will ensure that more digital resources are available to support the use of Welsh" (Welsh Government, 2018).

This plan identified three specific areas to be addressed, namely:

1. Speech technology
2. Computer-assisted translation
3. Conversational Artificial Intelligence.

In each case the challenges are addressed, and in addition, underpinning themes are elaborated, including:

- Creating and sustaining digital infrastructure
- Developing a culture of open innovation
- Building capacity and digital skills

- Digital transformation in the public sector
- Promoting the creation and use of Welsh language digital products and services.

Welsh is therefore in a privileged position for a minoritized language in having a well-defined roadmap for future action.

3. Relevant Projects So Far

Several projects have been undertaken in recent years to lay the foundations for a coherent programme of language technology development for Welsh. Even without large scale, long term funding, academic researchers have been able to refer to government policies when making grant applications for the short-term projects then on offer. Previously there was no strong tradition of research in speech technology, machine translation or AI in Welsh universities, although it can be argued that these were in any case new fields of study, and that developing these areas for Welsh meant that Wales gained important new capacity in these fields. The main research so far has been conducted at Bangor University in north Wales, with some related activity in other institutions, and renewed efforts to establish an all-Wales research network in language technologies to advance the field in general.

Early projects concentrated on text-to-speech, since visually impaired Welsh speakers in the early 2000s were unable to access e-mails, text documents, and other materials on their computers that was written in Welsh. The ground-breaking WISPR (Welsh and Irish Speech Processing Resources) project was funded by the Interreg IIIA EU programme, leading to the first easy-to- use synthetic voices for Welsh, and later also for Irish (Williams, Prys and Ní Chasaide, 2005). Later research developed speech recognition resources for Welsh (Cooper, Jones and Prys, 2014; Prys and Jones, 2018 (1)), this time funded by the Welsh Government. Resources first developed in the WISPR project for text-to-speech, such as a Welsh pronunciation lexicon, were reused and updated, as part of a philosophy of making the best use of resources available.

In the meantime, advances in machine translation (MT) was bringing ever improving results for English and some other major languages. Minoritized languages often exist in bilingual environments alongside the dominant major language, and local translation industries have developed as a result. A report on translation tools for the translation industry in Wales (Prys, Prys and Jones, 2009) discussed MT tools for Welsh, leading to further research on MT for the Welsh-English language pair. Importantly, a Knowledge Transfer Partnership (KTP) project with a local translation company enabled that company to develop high quality domain specific MT using the company's own vast archive of legacy translations (Prys and Jones, 2019). Again, it was possible to share and reuse some resources such as wordlists with the speech technology projects.

The most recent addition to this mix of language technologies has been conversational artificial intelligence. As part of the 'Macsen' project to create a prototype personal assistant in Welsh (Jones and Cooper, 2016), spoken questions had to be understood and replied to appropriately. Although at a basic level this was possible by listening for some key words and using various APIs to provide answers relating to news, weather and time, to progress further in this field research is needed in intent parsing, natural language generation and many other new areas. This may seem overly ambitious for a small language, but in bilingual communities, where public authorities and private companies are moving towards AI conversational agents for reasons of cost and efficiency, the minoritized language has no option but to try and keep up. Again, reusing existing datasets and resources, refining them, and donating them back to the community go some way towards making such projects achievable.

The open-source platforms used include MaryTTS (Pammi et. al. 2010) for text to speech, Kaldi (Povey et. al. 2011) and later DeepSpeech (Mozilla) for speech recognition, and Moses-SMT (Koehn et. al. 2007) for MT. However, some of these platforms are challenging for others to use, and containerized wrappers have been developed at Bangor University for Moses and DeepSpeech (Jones, 2015 and 2018) to make this software more user friendly for non-experts. These platforms are useful for a large number of languages and contribute greatly to keeping down costs in developing LTs for less-resourced languages.

4. Licencing and Dissemination

The release of data itself under open source licences has been the subject of some debate in Wales as elsewhere. Large datasets are one of the core requirements to train any models in speech technology, MT and conversational AI agent applications. Finding enough appropriately licenced data is one of the biggest challenges for less-resourced communities. Any strategy for efficient and effective harvesting of data can make the difference between supporting a language or not in a software package. For example, attempts have been made with apps specifically developed to crowdsource a speech corpus from the Welsh language community (Cooper, Jones and Prys, 2019). More recent activity in crowdsourcing Welsh language speech data has focused on collaborating with and sharing efforts with Mozilla's CommonVoice initiative, since its philosophy and motives align (Prys and Jones (2018 (1)). Not all language communities are happy to lose control of their data, especially where they have had bad experiences of colonial exploitation in the past. However, in the Welsh context, the use of crowdsourcing strategies and of permissive licensing of data has helped the development of Welsh language software by the private sector, aided in some cases by knowledge transfer partnerships between academia and industry.

Even in Wales it has not always been possible to release data on open source licence, as some legacy products came with their own, previous licences. In other cases, there was the need to sell commercial software in order to fund the continuation of the work. Increasingly however, and wherever public funding was used to create the tools and resources, they were released on permissive licences such as BSD, MIT, Apache or CC-0, which permit reuse without any restrictions. This was in order to make the tools and

resources attractive to enable both small and large companies to take up and use in their own products. In both cases, the private sector is less willing to take up tools and resources published under copyleft licences, such as GPL and CC-BY-SA, which stipulate that the entire utilising body of software must be released openly under the same licence.

Although Welsh, with its approximately half a million speakers is deemed to be a very small market for commercial companies, it is still large enough to support many small companies who could benefit from language technology tools and resources. These include translation companies, local media, software companies, web designers, and producers of educational games and language teaching materials. In common with the experience of many other minoritized and endangered languages in peripheral regions, there are high proportions of Welsh speakers in the rural and remote north and west of Wales, areas that are impoverished with few opportunities for well-paid employment and therefore suffer from emigration of young, talented people. Providing appropriately licensed language resources to small, local companies in these areas can therefore help make them viable and help economic as well as linguistic revitalization of these areas.

The arguments for releasing resources on permissive licences for large multinational companies are somewhat different. It can be argued that multinationals can well afford the development costs of including smaller languages amongst their multilingual offerings, and that paying for the necessary linguistic resources would be a great help to those languages. However, in the absence of strong legislation requiring Welsh language provision, most multinationals only heed the economic argument, and if the cost of producing or procuring those resources is larger than the anticipated return on their outlay, they will not pay for their development. If, on the other hand, appropriate resources are available to them at no cost, they are then more willing to consider supporting that language amongst their offerings. The minoritized or endangered language community benefits as many of their users already use those products every day in English, Spanish, French or whatever other dominant language they speak.

Additional clarity would however be welcomed in understanding the legal ramifications of different licences as there are many legal grey areas. For example, if new language or acoustic models are trained from a specific corpus, does the licence of the original corpus carry over to the new models? Or when a new MT engine is trained on a certain dataset, how does that affect the licencing of the new product? This is especially problematic when there are different licences for the two languages in a bilingual corpus, especially derived from a translation memory where the copyright of the original language text was not originally made explicit.

If tools and resources are to be shared outside individual projects and institutions, then dissemination is another issue that comes to the fore. International repositories such as Metashare, Github and Docker Hub have made it easier for developers to find resources in different languages, but for the non-expert user, and anyone interested in a specific language, the plethora of different repositories can be confusing. In addition to using the international repositories therefore, a Welsh National Language Technology Portal was established as a ‘one stop shop’ or ‘brochure site’ pointing at the different resources and giving additional guidance and information on their use (Prys and Jones, 2018).

5. Next Steps

When the Welsh government published its Language Technologies Action Plan in 2018, we could see from the account above that some preparatory research and development had already been done. Work had already begun on the three main areas to be addressed: speech technology, machine translation and conversational AI. Further long-term funding is likely to progress in these areas, with a coherent action plan providing further guidance. The five underpinning themes mentioned in the action plan (creating and sustaining digital infrastructure, developing a culture of open innovation, building capacity and digital skills, digital transformation in the public sector and promoting the creation and use of Welsh language digital products and services) are also challenges to be faced. The themes of course are broad in scope and will need cooperation from many different stakeholders. They demonstrate that developing language technologies and using them to revitalize a language cannot happen in isolation from the wider infrastructural and cultural environment.

The research base in Wales remains very small and issues such as building capacity and digital skills need urgent attention. This is being addressed for school children in the new 2022 National Curriculum for Wales (A Guide to Curricul for Wales, 2019), with its emphasis on digital competence as one of the three core essentials (the other two are literacy and numeracy). At the other end of the educational journey there are plans to create a new Masters in Language Technology programme at Bangor University. Delivering new tools and resources for use in commercial products and services is also crucial. There are opportunities here for the emerging creative and software sectors in Wales, with the potential to develop a new domestic market and to venture into wider multilingual markets from a strong bilingual base.

Some of these points were also echoed in a roundtable discussion to promote a strategic vision for Celtic Language Technologies held during the Celtic Congress at Bangor (Prys and Williams, 2019). Many participants were members of the Celtic Language Technologies Group (CLT), a lose grouping of academics who encourage research in language technologies for the various Celtic languages and organise occasional workshops in the field. All six of the modern Celtic languages are minoritized and endangered, with only Welsh and Irish having developed any coherent strategies for using language technologies for language revitalization. During the roundtable discussion the main needs were summarized as follows:

- the sharing of information across researchers working on individual Celtic languages

- working together to improve training and providing courses in Language Technologies
- developing transfer learning methodologies and common language models for our languages
- sustainability and long-term solutions to maintain our resources.

Training and sustainability needs were themes also picked up in the Welsh Government Action Plan and are doubtless also relevant to other minoritized language situations. Improving the sharing of information amongst the wider community and developing joint research proposals for transfer learning and common models are action points for the CLT to take forward. An official planning document on language technology development for the Irish language is also eagerly awaited, and together demonstrate concerted efforts to allow language technologies to contribute significantly to language revitalization, at least in Wales and in Ireland.

6. Conclusions

Welsh is in a fortunate position compared to many other minoritized and endangered languages. The groundwork has been laid for further development of language technologies, and the various stakeholders: academic researchers, public bodies and industry, are poised to take advantage of new opportunities offered by the Welsh Government's Action Plan. Most importantly, the language community itself is engaged, both as consumers of digital materials and devices and as potential producers of new Welsh digital content and software. Language technologies offer a path towards economic regeneration as well as language revitalization, and while there is much hard work still to be done, current progress is encouraging.

7. Acknowledgements

We gratefully acknowledge funding and support from the Welsh Government for the development of Welsh speech and language technologies since 2013.

8. References

A Guide to Curriculum for Wales 2022. 2019. Welsh Government. https://hwb.gov.wales/storage/f8f9760c-64a1-48ea-80fd-db130ad9050b/a-guide-to-curriculum-for-wales-2022.pdf [Accessed: 12 January 2020].

Cooper, S., Jones, D.B. and Prys, D. (2014). Developing further speech recognition resources for Welsh. In John Judge et al., editors, *Proceedings of the First Celtic Language Technology Workshop at the 25th International Conference on Computational Linguistics (COLING 2014).* Dublin, Ireland. Pages 55-59.

Cooper, S; Jones, D.B & Prys, D. (2019) Crowdsourcing the Paldaruo Speech Corpus of Welsh for Speech Technology. In a special edition on Computational Linguistics for Low Resource Languages, *Information*, 2019. 10, 247.

Jones, D.B. (2015). Docker Container for Moses. Available at https://hub.docker.com/r/techiaith/moses-smt [Accessed: 11 January 2020].

Jones, D.B. (2018). Docker Container for DeepSpeech. Available at https://hub.docker.com/r/techiaith/deepspeech [Accessed :11 January 2020].

Jones, D.B. and Cooper, S. (2016). Building Intelligent Personal Assistants for Speakers of a Lesser-Resourced Language. In Claudia Soria, et al., editors, CCURL Workshop. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. European Language Resource Association (ELRA). Pages 74-79.

Koehn, P.; Hoang, H.; Birch, A.; Callison-Burch, C.; Federico, M.; Bertoldi, N.; Cowan, B.; Shen, W.; Moran, C.; Zens, R.; Dyer, C.; Bojar, O.; Constantin, A. and Herbst, E. (2007), Moses: Open Source Toolkit for Statistical Machine Translation. In John A. Carroll; Antal van den Bosch & Annie Zaenen, editors. *'ACL'*, The Association for Computational Linguistics. Pages 177-180.

Mozilla (n.d.). A TensorFlow implementation of Baidu's DeepSpeech architecture. Available at https://github.com/mozilla/DeepSpeech [Accessed: 11 January 2020].

Office of National Statistics. (2012). *Language in England and Wales: 2011.*

Pammi, S., Charfuelan, M., and Schröder, M. (2010). *Multilingual voice creation toolkit for the MARY TTS platform.* In 7th International Conference on Language Re-sources and Evaluation (LREC), Valletta, Malta. Pages 3750–3756.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlíček, P., Qian,Y., Schwarz, P., Silovský, J., Stemmer, G., and Veselý,K. (2011). The Kaldi speech recognition toolkit. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU).* Waikoloa, HI, USA.

Prys, D. and Jones, D.B. (2018 (1)). Gathering Data for Speech Technology in the Welsh Language: A Case Study. In Claudia Soria, et al., editors, CCURL Workshop. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC'18).* European Language Resource Association (ELRA). Pages 56-61.

Prys, D. and Jones, D.B. (2018 (2)). Language Technologies Portals for LRLs: A Case Study. *Lecture Notes in Artificial Intelligence.* Springer. Pages 420-429.

Prys, D., Prys, G., and Jones, D.B. (2009). *Improved Translation Tools for the Translation Industry in Wales: An Investigation.* Bangor University, Bangor, Wales.

Prys, M. and Jones, D.B. (2019). Embedding English o Welsh MT in a Private Company. In Teressa Lynn et al., editors. *Proceedings of the Celtic Language Technology Workshop.* European Association for Machine Translation. Dublin, Ireland. Pages 41-47.

Prys, D. and Williams, I. (2019). *A Round Table Discussion to Promote a Strategic Vision for Celtic Language Technologies.* Bangor University, Bangor. Available at http://techiaith.bangor.ac.uk/wp-content/uploads/2019/08/A-roundtable-discussion-to-promote-a-strategic-vision-for-Celtic-Language-Technologies.pdf [Accessed: 11 January 2020].

Williams, B., Prys, D. and Ní Chasaide, A. (2005). Experiences of creating a research capability in speech technology for two minority languages. In *Proceeding of the 9th European Conference on Speech Science and Technology (Interspeech).* Pages 188-191.

Proceedings, pages 371–374
Paris, UNESCO Headquarters, 5-6 December, 2019.

Archiving System of Endangered Languages in Japan: A Preliminary Report

Natsuko Nakagawa, Masahiro Yamada, Nobuko Kibe, Yukinori Takubo
National Institute for Japanese Language and Linguistics
10-2 Midoricho, Tachikawa, Tokyo, Japan
{nakagawanatuko, m-yamada, nkibe, ytakubo}@ninjal.ac.jp

Abstract

There are eight (UNESCO), twelve (Ethnologue), or more (degree of intelligibility) endangered languages/dialects in Japan. We present a database and digital archiving space that NINJAL (National Institute for Japanese Language and Linguistics) is developing for all of these languages where individual researchers or language communities can deposit their field data, language documentation, or audio-visual recordings. Two major features of the database/archiving space include (i) that it is a Japanese-mediated database/archive and thus virtually everyone in Japan can use it, and (ii) that it comes with an online exhibition space so that archiving is tightly connected to public use of the deposited items.

日本には 8（ユネスコ）、12（エスノローグ）、もしくはそれ以上（相互理解性）の消滅の危機に瀕した言語・方言が存在する。本発表は国立国語研究所がこれらの言語・方言のために開発中の、個別の研究者や言語コミュニティが利用可能なデータベースおよび電子的アーカイブスペースについて報告する。データベース・アーカイブスペースは以下の二つの特徴を持つ。(i) 日本語によるデータベース・アーカイブスペースであり、日本に住む誰もが利用可能である。(ii) オンライン展示スペースが付随し、アーカイブされるデータが社会一般に対する公開と密接に結びついている。

Keywords: digital archiving, exhibition, collaboration with communities, languages in Japan

1. Introduction

Atlas of the World's Languages in Danger (Moseley, 2010) reports that there are eight endangered languages in Japan: Ainu, and seven Japonic languages, namely Hachijo, and six Ryukyuan languages (Figure 1.). Although these eight languages tend to receive a special attention with respect to language conservation, there are many more mutually unintelligible sub-varieties including Japanese dialects. Indeed, there have been eight reference grammars for Ryukyuan and three for Japanese dialects, and many more grammar sketches (Shimoji, 2008; Pellard, 2009; Niinaga, 2014; Shimoji and Pellard, 2010; Heinrich et al., 2015). Under the influence of the dominant language, so-called Standard Japanese, most, if not all, of such local languages are in danger of extinction.

Since 2010, National Institute for Japanese Language and Linguistics (NINJAL) has been leading a collaborative research project termed the *Kiki-Gengo Project* (Endangered Language Project, PI: Nobuko Kibe) to describe and conserve the local languages in Japan. As of 2019, 57 researchers from 36 domestic and overseas universities/institutions are working on documentation of the local languages spoken at 40 sites across Japan and producing vocabulary lists, grammar sketches, and narratives. The project also collaborates with some of the local governments to raise the awareness of the language endangerment.

In this paper, we will first present the currently available language resources resulting from the NINJAL project. We also introduce an ongoing project of language archiving and an exhibition system that targets researchers and language communities.

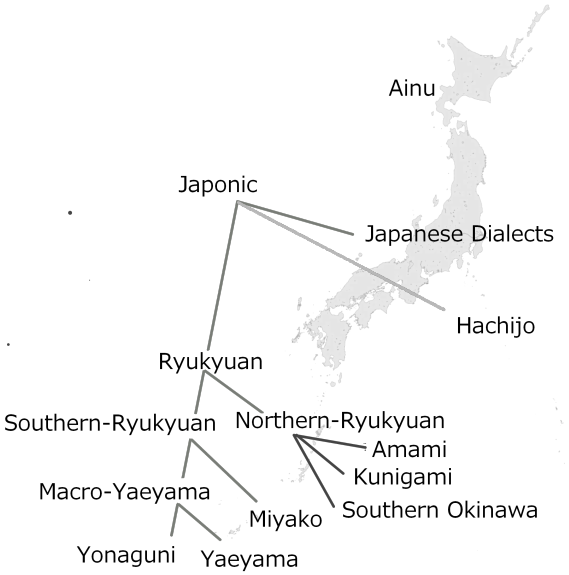


Figure 1: Language map in Japan

2. Current Resources

The documentation works done by the collaborative researchers in NINJAL's *Kiki-Gengo Project* are incrementally made available at *Kiki-Gengo Database* (Endangered Language Database). NINJAL also provides a corpus of Japanese dialects (COJADS). This section gives a brief overview of these resources. Also, some of the colleagues made a set of fonts called "Shima Shotai Font" in collaboration with a designer so that everyone can type their languages on computer. We give an overview of these re-

sources.

2.1. Endangered Language Database

The *Kiki-Gengo* (endangered language) Database is a website to make NINJAL's collaborative researchers' documentation works accessible online (Celik and Kibe, 2019).[1] It provides vocabularies and narratives site-by-site and currently covers 14 sites (mainly Hachijo and the Ryukyus). Approximately 9,000 words and expressions are currently registered and searchable online. In addition to the recordings collected by the collaborative researchers, some legacy recordings of dialogues and monologues collected as old as 1969 are also available.

Grammar sketches will be made available at the website by the end of the current project period, 2021. For now we post pdf versions of the grammar sketches along with a basic vocabularies and sentences at nine sites.

Currently the data depositing is done by a few project RAs by hand and is not quite functioning as an ideal archiving system, but the process is fairly simple for the collaborative researchers for they just need to submit their field data filled in a spreadsheet template along with the audio files.

2.2. Corpus

Corpus of Japanese Dialects, or COJADS is the largest corpus of Japanese dialects currently available (Aoyama et al., 2019; Kibe et al., to appear). Currently, the beta version is available with the data from 48 sites (two from Okinawa prefecture and one for each of the 46 prefectures of Japan); 30 minutes of narratives for each site, for a total of 24 hours (published as NINJAL (2001-2008)).

The corpus is searchable both in transcriptions and in Standard Japanese translations at NINJAL's Chunagon corpus concordance system.[2] The audio data corresponding to the search result is also presented on the same page.

The original data used on COJADS consists of more than 4,000 hours of narratives collected from 224 locations through the "Urgent Survey of Japan's Regional Dialects", a nationwide investigation project conducted between 1977 and 1985 by the Agency for Cultural Affairs. COJADS aims to cover 75 hours of narratives by 2021.

ぼーや　んめー、　んめ　すでぃりてぃーぬ　（ばなす）。
かぬ、　すでぃみずぅー　かたみー、　てぃんから　たすき、
つかいぬ　みどぅむぬ　いーぬ　ぬくりーりてぃーどぅ、…。
えー

Figure 2: Shima Shotai Font

2.3. Shima Shotai Font

Many languages/dialects in Japan have sounds that cannot be written in Japanese writing, which is based on the Standard Japanese sound system. Ogawa (2015) proposed a unified writing system to express all sounds in Ryukyuan varieties. The proposed system, however, includes unconventionalized usage of diacritics (such as ˚ , ˋ , and ^) and needs a new type of font. Ogawa et al. (2019) proposed new font termed "Shima Shotai Font" (*shima* means "(our) island" and "(our) village" across Ryukyuan islands), which includes unconventionalized combinations of diacritics and kana (Japanese phonogramic characters). See Figure 2 for an example. It is *kana*-based phonogramic system instead of *kanji*-based ideogramic system because the pronunciation of *kanji* is often ambiguous. The set of fonts makes it easier for communities to use their languages on PCs (and smartphones in the future), which is expected to promote their language usage. Vertical writing is also possible with this set of fonts. In the future, Ogawa and his colleagues want the font to be included in Unicode fonts.

3. Archiving and Exhibition Space

This section discusses our ongoing project and future plan.

3.1. Language Archiving System

We are now building a language archiving system both for researchers and for community members. The system should be Japanese-mediated so that almost all citizens in Japan can access it. It should be easily depositable and immediately visible/usable; i.e., it should be easy for everyone to deposit their data and to see and search the data immediately after they deposit them.

3.2. Language Museum

The language archiving system accompanies with exhibition space, called "digital museum of language", which exhibits part of the archived data mainly for non-researchers (Takubo, 2017). This kind of exhibition space not only encourages the community members to be proud of their community, but also motivates researchers to deposit more data since researchers receive feedback from the community. The goal is to receive positive feedback, but should the feedback be negative, researchers would strive to improve the content.

We plan to show our materials with coherent stories. For example, users interested in the history of Miyako Island will be able to access related videos in chronological order, while users interested in sightseeing can access videos introducing popular locales and other nearby attractions. Multiple stories will be attached to the same video, accessible based on the users' interests. In the future, we would like to make it possible for everyone to make their own stories, arrange materials in the museum, and publish them.

Materials for Miyako language (Ikema-Nishihara variety) learning are already partially available.[3] They cover pronunciation, the writing system, grammar, and videos. We will improve this website to be more user-friendly.

4. For and with Communities

This section introduces some projects we are working on with community members.

4.1. Picture books with audio, video, and text

We and other colleagues (Akiko Tokunaga-Yokoyama, Kayoko Shimoji, and Fumi Yamamoto) are making four

[1] http://kikigengo.ninjal.ac.jp

[2] https://chunagon.ninjal.ac.jp

[3] http://www.kikigengo.jp/nishihara/doku.php

Figure 3: Drawings from a picture book

picture books for children in four different Ryukyuan languages. They are based on the local stories or songs inherited from their ancestors. The pictures were drawn by Fumi Yamamoto, a professional illustrator and designer, who happened to be attracted to one of the stories in the Ryukyus (see Figure 3).

Each story was narrated by a local native speaker, was transcribed by a linguist who specializes in the language, and was accompanied with interlinear gloss and translations (in Japanese).[4] The linguists also wrote grammar sketches to help readers to read aloud the text and understand the story in the given language (Nakagawa and Yamada, 2018). Narration by the native speaker was recorded and will be available online so that people can learn how to pronounce each word and how to read a sentence with natural intonation. The books have a kid-friendly appearance and content, but it is also intended that the parent generation can practice the language while they read them to their children. The parent generation are considered to be passive speakers; they can understand the language but do not speak it (Yamada et al., to appear). The picture books are expected to encourage them to speak their language.

[4]English translations will be also available.

We are now raising a fund via crowdfunding to publish the books.[5] 213 people donated 3,032,000 Japanese yen (approximately 28,000 US dollars) as of 13th January, 2020. Comments left by the supporters tell that the supporters are not necessarily related to Ryukyu.

4.2. Dictionary

Traditionally, a single native speaker devotes their life to compiling a dictionary, and it usually takes decades to publish it. Now, younger community members collect words and expressions from elderly speakers with a help of linguists and input the collected items in Google Docs. Another linguist in Tokyo receives the data immediately, change the spread sheet into a database and makde it available online. Another linguist prints the data in a dictionary-like format (using LaTeX), and send it back to the community. In the future, we want to make an online dictionary maker which automatically generates a professional dictionary from Google Docs (or tab-separated file).

4.3. Interactive interlinear glossed transcriptions (Karaoke)

We are making a system of interactive interlinear glossed transcript called *karaoke*. For now, only karaoke system with sounds (without video) is available.[6] In this system, users can not only see the subtitles while watching a video or listening to narratives, but also can click the subtitles and watch and listen to the corresponding part. Users can also chose what they see; they can watch a video or listen to narratives without any subtitles or they can chose to see one of the transcriptions (IPA, *kana*-based style, and/or only translations). In this way, the karaoke system functions as an efficient, customized learning tool. It can also be applied to songs.

4.4. Physical “escape the room” game

We (Niinaga Yuto, Masahiro Yamada, and Natsuko Nakagawa) are also making a game called physical “escape the room” game with Igengo Lab. “Escape the room” is a kind of game played with a book or, more recently, a computer. The player is confined to a room and is required to escape from the room by collecting hints hidden in the room. The hints are often encrypted, and the player needs to infer the solution. In the physical escape the room game, the players actually play the game physically. They (usually a team) are confined in a room, seek hints often encrypted, and look for the solution.

Igengo Lab is one of such game-developping groups. They specialize in Japanese Sign Language and have made various problems (e.g., encrypted hints) that can be solved only when signers and non-signers collaborate with each other.

We and Igengo Lab made a game using three varieties of Ryukyuan: Yonaguni Dunan, Shiraho (Ishigaki, Yaeyama), and Amami (See also Figure 1.). The players are required to compare these three languages, find phonological correspondence and some grammatical rules, and communicate with local people in the local language in order to solve the problems. We will run the first trial in February 2020 in

[5]https://readyfor.jp/projects/minato

[6]http://kikigengo.ninjal.ac.jp/danwashiryo.html

Tokyo and hope to continue in the Ryukyu Islands in the future.

5. Conclusion

This paper gave a brief overview of NINJAL projects related to endangered languages in Japan. The resources already available are a vocabulary database (9000 expressions from 14 places, contemporary), spoken corpora (75 hours from 25 places, approximately 40 years ago), and the Shima Shotai Font, expected to be more ubiquitous and be registered as Unicode fonts. We are now building an archiving system with exhibition spaces that are mediated in Japanese. Finally, we introduced four ongoing projects with communities: picture books, dictionaries, a karaoke system, and a physical "escape the room" game.

6. Acknowledgements

The work reported in this article was supported by JSPS KAKENHI projects 16H01933, B16K16824, and 18K12360.

7. Bibliographical References

Aoyama, K., Asahara, M., Carlino, S., Ishimoto, Y., Kibe, N., Koiso, H., Maekawa, K., Nishikawa, K., Wakasa, A., Watanabe, M., and Yoshikawa, Y. (2019). Speech corpora in NINJAL, Japan: Demonstration of corpus concordance systems: Chunagon and Kotonoha. In *The 3rd International Symposium on Linguistic Patterns in Spontaneous Speech*, Taiwan. Institute of Linguistics, Academia Sinica.

Celik, K. and Kibe, N. (2019). Raising language diversity awareness in Japan through web-based open access application. In *6th International Conference on Language Documentation and Conservation (ICLDC)*.

Patrick Heinrich, et al., editors. (2015). *Handbook of the Ryukyuan Languages*. Mouton de Gruyter, Berlin.

Kibe, N., Sato, K., Nakanishi, T., and Nakazawa, K. (to appear). Corpus-based study of japanese dialects: Regional differences in accusative case marking system. In *Proceedings of Methods in Dialectology XVI.*

Christopher Moseley, editor. (2010). *Atlas of the World's Languages in Danger*. UNESCO Publishing, Paris.

Nakagawa, N. and Yamada, M. (2018). Taketomi-jima hoshi-suna-no hanashi-no ehon seisaku-to ippan-dokusha-muke bumpô-gaiyô-no shippitsu [making the picture book a tale of star sand in taketomi and writing a grammar sketch for a general audience]. *NINJAL Research Papers*, 14:145–167.

Niinaga, Y. (2014). *A grammar of Yuwan, a Northern Ryukyuan language*. Ph.D. thesis, University of Tokyo.

Ogawa, S., Yamada, M., Hayashi, Y., and Ueda, H. (2019). A typeface for endangered languages. In *ATypI2019*.

Shinji Ogawa, editor. (2015). *Ryukyu-no Kotoba-no Kakikata [How to write Ryukyuan languages]*. Kurosio, Tokyo.

Pellard, T. (2009). *Ōgami – Éléments de Description d'un Parler du Sud des Ryūkyū*. Ph.D. thesis, École des hautes études en sciences sociales. (Ōgami – Elements of description of a Southern Ryukyuan language).

Michinori Shimoji et al., editors. (2010). *An introduction to Ryukyuan languages*. ILCAA, Tokyo.

Shimoji, M. (2008). *A Grammar of IRabu, a SOuthern RYukyuan Language*. Ph.D. thesis, Australian National University.

Takubo, Y. (2017). The digital museum project for the documentation of endangered languages: The case of Ikema Ryukyuan. In A. Vovin et al., editors, *Studies in Japanese and Korean historical and theoretical linguistics and beyond*, pages 3–12. Brill, Leiden.

Yamada, M., Takubo, Y., Iwasaki, S., Celik, K., Harada, S., Kibe, N., Lau, T., Nakagawa, N., Niinaga, Y., Otsuki, T., Sato, M., Shirata, R., Van der Lubbe, G., and Yokoyama, A. (to appear). Experimental study of inter-language and inter-generational intelligibility: Methodology and case studies of Ryukyuan languages. *Japanese/Korean Linguistics*, 26.

Proceedings, pages 375–378
Paris, UNESCO Headquarters, 5-6 December, 2019.

Efforts in the Development of an Augmented English–Nepali Parallel Corpus

Sharad Duwal, Bal Krishna Bal
Information and Language Processing Research Lab
Department of Computer Science and Engineering
Kathmandu University, Dhulikhel, Kavre, Nepal
sharad.duwal@gmail.com, bal@ku.edu.np

Abstract

A crucial resource for Machine Translation between any two languages is the amount of quality parallel data. High-resource language pairs have been abundantly studied, but this is not true in the case of under-resourced languages. However, the attention is now gradually shifting towards under-resourced languages as well. Efforts are underway for creating more parallel data. In this paper, we explain the procedures we followed to develop an augmented English–Nepali parallel corpus. We also report new baseline scores for the pair.

Keywords: corpus, parallel, synthetic, data pre-processing

सार संक्षेप

कुनै दुई भाषामा यान्त्रिक अनुवाद गर्दा गुणस्तरीय समानान्तर डेटाले अहम् भूमिका खेल्छ। यस सन्दर्भमा समानान्तर डेटाको दृष्टिकोणबाट उच्च स्रोतसम्पन्न भाषाका जोडीहरूमा जति शोध र अनुसन्धान भएको छ त्यति न्यून स्रोतसम्पन्न भाषाका जोडीमा भएको पाइँदैन। यद्यपि पछिल्लो समयमा विस्तारै मानिसहरूको ध्यान न्यून स्रोतसम्पन्न भाषातिर गएको पाइन्छ। यस्ता भाषाहरूमा बढीभन्दा बढी समानान्तर डेटाको सिर्जनाका लागि प्रयत्नहरू भइरहेको छ। यस शोध पत्रमा अङ्ग्रेजी नेपाली समानान्तर डेटाको संख्या बढाउन हामीले गरेका प्रयत्नका बारे हामीले व्याख्या गरेका छौँ।

1. Introduction

There has been limited research on Machine Translation (MT) for the English–Nepali language pair. The earliest works were based on lexicon and rules, namely the "Dobhase" project (Bista et al., 2005) and the Apertium-based English Nepali translation system (Dahal, 2011). Google Translate, a free multilingual MT service provided by Google, added Nepali language support in 2013. Google Translate follows a Neural Machine Translation (NMT) approach and provides relatively good translations in both English–Nepali and Nepali–English directions though not always of acceptable quality. Acharya and Bal (2018) conducted a comparative study of Statistical Machine Translation (SMT) and Neural Machine Translation (NMT) in which they concluded that SMT still underperforms NMT in case of low resourced languages. Given a substantial amount of parallel data, a low-resource language like Nepali can also benefit from NMT which is being looked into with great interest from researchers all over the world. In this paper, we explain the procedures we followed to develop an augmented English-Nepali parallel corpus. The result of developing such resource has been very encouraging in terms of increased BLEU scores compared to the baseline scores which we discuss in detail in the following sections.

1.1 Related Work

There have been very few documented efforts towards the development of a parallel corpus for the Nepali language. Yadava et al. (2008), as part of the Nepali National Corpus (NNC) project, collected around 13.8 million words of monolingual Nepali written corpus and around 4 million words of parallel data (English–Nepali). A more recent work done by Guzmán et al. (2019) involved the creation of several evaluation sets for the English–Nepali pair. They used sentences originating from the Wikipedia of both the languages and had them professionally translated. To ensure the quality of these translations, they used several automatic and manual filtering procedures.

Although sparsely located and not very well-documented, several sources contribute to the parallel data for the English–Nepali pair. These include the Linux translations under OPUS (Tiedemann, 2012), Bible translations[1], translations of the Penn Treebank under the PAN Localization Project[2], Global Voices 2018q4 corpus, etc. These amount to about 600k parallel sentences and yield BLEU scores of 7.6 (NE-EN) and 4.3 (EN-NE) in supervised setting on the *devtest* evaluation set prepared by Guzmán et al. (2019).

However, upon closer inspection, we found these parallel corpora to be quite noisy and poorly aligned. The major problem seemed to be misaligned and repeated sentences. We employed different techniques ranging from manual cleaning to automated processes to cull around 100k parallel sentences from the 600k sentences. We then collected around 110k additional parallel sentences and consolidated them all to form a semi-parallel corpus (mix of truly parallel and synthetic parallel data).

1.2 Corpus Details

The individual contribution of each source to the development of the augmented corpus has been shown in Table 1. We report the new sources and mention our efforts to clean the previously available data.

1. **GNOME/Ubuntu/KDE from OPUS**: In this corpus (Tiedemann, 2012), we found that sentences were heavily repeated. Using length-based comparisons between sentences and their translations, we

[1]https://github.com/christos-c/bible-corpus
[2]http://www.cle.org.pk/software/ling_resources/UrduNepaliEnglishParallelCorpus.htm

manually removed or edited misaligned sentences. At the end, we brought the corpus down from more than 500k parallel sentence pairs to just 60k sentence pairs.

2. **Bible Translation:** We used only one edition of the two Bibles available in English. We did not use the archaic English Bible. Misaligned sentences were few and far between and only a few of them required correction. The final contribution was about 30k parallel sentences.

3. **Global Voices 2018:** These were completely misaligned most of the time and a thorough manual cleaning was done.

4. **NNC Parallel Corpus :** Around 17k parallel sentence pairs were extracted from six documents of the NNC which were basically document-aligned. The documents were National Development Plan texts. We ran a few custom scripts to break the texts into sentences and used Bleualign (Sennrich and Volk, 2011) with automatic translations generated by a Nepali–English model trained on the corpora described so far (Bible, OPUS, Global Voices 2018). In the first iteration we obtained around 13k sentence. We trained a new model with the new data for a second iteration and obtained 16k sentences. The third iteration yielded 17k sentence pairs.

5. **Nepal Law Texts:** The Nepal Law Commission website[3] makes available laws, acts, policies, etc. in both Nepali and English. These were also document-aligned, so we applied a similar procedure to these as we did with the NNC parallel corpus documents (described above) and after two iterations obtained around 60k parallel sentences.

6. **Easy Bible:** We found a completely different version of Bible and its own Nepali translation[4] from which we obtained 31k sentence pairs after some manual processing.

7. **Synthetic Corpus:** We crawled popular English and Nepali newspapers and conducted a careful and highly-iterative backtranslation procedure (Sennrich et al., 2016a). We also used most of the monolingual Nepali corpus collected under NNC. At a given iteration, we augmented the corpus with as many sentences as it currently had and backtranslated for two iterations with the same amount of data and only then went to the next iteration. We retranslated the monolingual data from both the directions before adding new monolingual data. Synthetic corpus contributes to over 1.6 million parallel sentences, which upon random manual inspection, was found to be moderately good. Finally, we added 688,801 parallel sentence pairs from monolingual English and 914,105 parallel sentences pairs from monolingual Nepali.

	Source	**Final Contribution** (No. of sentences)
1	GNOME/KDE/Ubuntu (OPUS) (Tiedemann, 2012)	58,645
2	Bible Translation (`https://github.com/christos-c/bible-corpus`)	30,577
3	Global Voices Parallel Corpus 2018q4 (`http://casmacat.eu/corpus/global-voices.html`)	4,696
4	Penn Treebank + (Acharya and Bal, 2018)	6,963
5	NNC Parallel Corpus	16,662
6	Nepal Law Commission	59,747
7	Easy Bible	31,498
8	Synthetic Corpus	1,602,906
	Total	**1,811,694**

Table 1: Details of the corpus (final individual contributions, after cleaning)

The final corpus has 1811694 parallel sentences, as presented in the Table 1.

2. Experiments and Results

In this section we describe the settings we use for training the baseline models and report the results.

2.1 Training Settings

We first consider a fully supervised setting in which we use only the 208k true parallel sentences and do not use the synthetic corpus.

Second, we consider the entire corpus. This can be considered to be a semi-supervised setting since we include the synthetic corpus as well.

The backtranslations have been performed using beam search with beam width of 5 and length penalties between 1 and 1.2.

2.2 Baseline Architectures

For all our experiments, we use the Transformer model (Vaswani et al., 2017) as implemented in the Fairseq toolkit (Ott et al., 2019). In the supervised setting, we use a Transformer with 5 encoder and 5 decoder layers, word representations of size 512, feed-forward layers with inner dimensions 2048, and 8 attention heads. In the semi-supervised setting, we use a similar Transformer architecture with 6 encoder and decoder layers.

[3] http://www.lawcommission.gov.np/

[4] https://ebible.org/find/details.php?id=npi2010

	Nepali–English		English–Nepali	
	devtest	*Test*	*devtest*	*test*
Supervised				
Guzmán et al. (2019) (600k)	7.6	–	4.3	–
Our work (208k)	**13.0**	12.17	**4.97**	4.9
Semi-supervised				
Guzmán et al. (2019) (Supervised + 5M Backtranslation)	15.1	–	6.8	–
Our work (208k + 1.6M Synthetic)	**20.76**	19.37	**7.49**	7.6

Table 2: BLEU scores in the Nepali–English and English–Nepali directions on the *devtest* and *test* sets. Scores on *test* set are not available for the (Guzmán et al., 2019).

We set dropout, weight decay and label smoothing to be 0.3, 10^{-4} and 0.1 respectively. We optimize the models with Adam optimizer (Kingma and Ba, 2015) with betas 0.9 and 0.98. For the smaller models we use a learning rate of 10^{-3} and for the larger models we use a learning rate of $7x10^{-4}$.

The final training sessions were carried out on a single NVIDIA Tesla P100 PCI-E GPU while many of the backtranslation iterations were carried out on a single NVIDIA Tesla V100 GPU.

2.3 Data preprocessing

1. **Tokenization:** We use the Nepali tokenizer in the Indic NLP library[5] to tokenize Nepali and we use the Sacremoses library[6] to tokenize English.
2. **Vocabulary:** We use the sentencepiece library[7] to learn joint Byte-Pair Encoding (BPE) (Sennrich et al., 2016b) on the source and target languages. We use vocabularies of sizes 2500 and 10000 for the supervised and semi-supervised settings respectively. We tokenize both the languages before learning the joint BPE over them.

We report scores on the *devtest* and *test* evaluation sets developed by (Guzmán et al., 2019). We report detokenized SacreBLEU (Post, 2018) when translating into English and tokenized BLEU (Papineni et al., 2002) when translating into Nepali.

2.1 Results

In the supervised setting, the BLEU scores on the *test* set and the *devtest* set are respectively 12.17 and 13.0 in the NE-EN direction and 4.9 and 4.97 in the EN-NE direction. In the semi-supervised setting, the BLEU scores (in the same order) are respectively 19.37 and 20.76 in the NE-EN direction and 7.60 and 7.49 in the EN-NE direction.

Against the baseline scores of 7.6 (NE-EN) and 4.3 (EN-NE) in the supervised setting reported by Guzmán et al. (2019), our NE-EN model scores almost double but the EN-NE model is only marginally better. In the semi-supervised setting, they ran two iterations of backtranslation on 5 million monolingual sentences and obtained BLEU scores of 15.1 (NE-EN) and 6.8 (EN-NE). Our models improve upon these baselines with much less data.

[5] https://github.com/anoopkunchukuttan/indic_nlp_library
[6] https://github.com/alvations/sacremoses
[7] https://github.com/google/sentencepiece

3. Availability

We plan to make the augmented corpus available for public use via the Information and Language Processing Research Lab, Kathmandu University very soon.

4. Conclusion

In this paper, we tried to address the major issue faced by under-resourced languages like Nepali—which is primarily the lack of quality parallel corpus. Developing a sentence-aligned quality parallel corpus is for sure a daunting and tedious task. Rather than developing something from scratch, we employed the strategy of leveraging on the existing sparse resources and at the same time augmenting them using different techniques like backtranslation of monolingual data. For a low resource language pair like Nepali–English, a parallel corpus of size 1.8 million sentence pairs is indeed a very valuable resource. This has been verified by the improvements achieved over the baseline scores. We believe the creation of this parallel corpus would trigger further research in the field of Nepali MT.

5. Acknowledgements

We would like to acknowledge our sincere thanks to Language Technology Kendra Nepal for making available the Nepali National Corpus (NNC). Our deep appreciation also goes to all contributors of the parallel corpus for the Nepali language.

6. Bibliographical References

Acharya, P. and Bal, B. K. (2018). A Comparative Study of SMT and NMT: Case Study of English-Nepali Language Pair. In *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*, pages 90–93.

Bista, S. K., Keshari, B., Bhatta, J., and Parajuli, K. (2005). Dobhase: online English to Nepali machine translation system. In *Proceedings of the 26th Annual Conference of the Linguistic Society of Nepal*, December.

Dahal, A. R. (2011). Development of a Nepali-English MT system using the Apertium MT platform. Technical report, The Language Technology Kendra, July.

Guzmán, F., Chen, P.-J., Ott, M., Pino, J., Lample, G., Koehn, P., Chaudhary, V., and Ranzato, M. (2019). The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical*

Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 6100–6113, Hong Kong, China, November. Association for Computational Linguistics.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, October. Association for Computational Linguistics.

Sennrich, R. and Volk, M. (2011). Iterative, MT-based sentence alignment of parallel texts. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, pages 175–182, Riga, Latvia, May. Northern European Association for Language Technology (NEALT).

Sennrich, R., Haddow, B., and Birch, A. (2016a). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany, August. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016b). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August. Association for Computational Linguistics.

Tiedemann, J. (2012). Parallel data, tools and inter-faces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.

Yadava, Y. P., Hardie, A., Lohani, R. R., Regmi, B. N., Gurung, S., Gurung, A., McEnery, T., Allwood, J., and Hall, P. (2008). Construction and annotation of a corpus of contemporary Nepali. *Corpora Vol. 3*, pages 213–225.

Proceedings, pages 379–383
Paris, UNESCO Headquarters, 5-6 December, 2019.

Rich Morphology, No Corpus – And We Still Made It. The Sámi Experience

Sjur Nørstebø Moshagen, Trond Trosterud
UiT The Arctic University of Norway
Department of Language and Culture
{sjur.n.moshagen, trond.trosterud}@uit.no

Abstract

The article presents an infrastructure for building grammar models and language technology applications for indigenous languages, i.e. for languages with a too complex grammatical structure and a lack of the huge amount of corpus material being necessary for mainstream language technology approaches to work. The infrastructure and grammar models provide a large array of applications for indigenous languages. The main problem for indigenous languages turns out to be structural hurdles set up by the major software providers. The article presents these hurdles, as well as a way of overcoming them.

Keywords: language technology, indigenous languages, morphology-rich languages, integration, application programming interfaces, localisation, language resource policies, standards

Čoahkkáigeassu (in North Saami)

Artihkal čájeha infrastruktuvrra, mii hukse grammatihkkamálliid ja giellateknologalaš prográmmaid eamiálbmotgielaide dahje gielaide, main lea kompleaksa giellaoahpalaš struktuvra. Eamialbmotgielain váilot maid teakstačoakkáldagat, mat leat dárbbašlaččat váldogielaid giellaoahpalaš reaidduid huksemii. Infrastruktuvrra ja giellamálliid bokte lea vejolaš hukset máŋggaid prográmmaid dávviguovlluid eamiálbmotgielaide. Váldováttisvuohta dál lea, ahte stuorra prográmmafitnodagat leat huksen prográmmaideaset nu, ahte ii leat vejolaš lasihit eamiálbmotgielaid heivehemiid fitnodagaid prográmmaide. Artihkal čájeha makkár váttisvuođaid birra lea sáhka, ja mo daid sáhttá čoavdit.

1. Introduction

This article presents a way of making grammatical analysers and language technology tools for minority languages with a complex grammatical structure. The infrastructure is hosted at The Arctic University.

The article is organised as follows: First we present our modus operandi, philosophy, scope of work and our major results. The next section presents the major problems facing language technology for minority languages, where integration of our solutions in major writing tools is a key point. The following section presents a solution to these problems, in the form of a manifesto for open language technology. Finally comes a conclusion.

2. Our modus operandi

We work together with the language community, and use rule-based technology and traditional linguistic analysis. In principle, one native speaker is enough, although we always try to involve the people or organisations that have the confidence of or are appointed by the language community to handle language planning.

Our models are not trained on any corpus, and the data sparsity bottleneck thus does not exist. Our technology supports the most complex morphology and morphophonology, but also works well for less complex languages.

By being grammar-based, we are able to offer a three-part cooperation, where all parties have something to gain: Participating in a project building language models within our infrastructure, university-based linguists will get the opportunity to test their grammatical generalisations on a full scale. Language activists will get the tools they need in order to revitalise the language, and programmers will get the opportunity to apply their programs on a new language. Opening up for a cooperation with university linguists is of crucial importance: Each language with a formalised orthography has at least one linguist devoting his or her life to this language. A grammar-based approach will offer this linguist good reasons for joining the project.

2.1. Open source, rule-based technology

In order to model the complex grammar of indigenous languages, we use the Helsinki Finite State Technology, Hfst[1], a tool set being source-code compatible with (Beesley and Karttunen, 2003), and ultimately going back to (Koskenniemi, 1983). It is capable of modeling natural languages in all its complexity, in an explicit and transparent way. To disambiguate morphological homonymy and adding syntactic function and thematic roles, we use Constraint Grammar ((Karlsson, 1990), (Karlsson et al., 1995)) and the compiler VISLCG3 [2].

The net result of this is a set of powerful grammatical analysers and generators for circumpolar and other languages[3]. In addition to providing a source of information on the grammar of these languages, the models are integrated in a language-independent infrastruc-

[1] `http://github.com/hfst`

[2] `http://visl.sdu.dk/cg3.html`

[3] An interface for using these programs can be found at `http://giellatekno.uit.no`

ture ((Moshagen et al., 2013)). By means of this infrastructure we are able to build a wide range of applications and tools. This includes a set of text proofing tools[4], programs for language learning[5] and for machine translation[6].

2.2. Mainly circumpolar languages

We primarily work with Saami languages, but also with other Uralic languages, with Greenlandic and Faroese, and with First nation languages in Canada (see e.g. (Arppe et al., 2016). All but Faroese are known for very to extremely rich morphology.

Figure 1: Morphology-enabled dictionaries in the present infrastructure

2.3. Lots of tools

Using these language models, we build spelling checkers, grammar checkers, mobile keyboards with spellers, rule-based machine translation, and offer analysed (i.e. machine-annotated) corpora[7]. We also make intelligent computer-assisted language Learning tools and dictionaries with grammars (cf. (Johnson et al., 2013), an overview of the languages is given in Figure 1).

Both the infrastructure, the linguistic resources, and most of the corpus texts are available as open source. Some corpus texts are not open, due to copyright issues, but single quotes are available via the corpus interface.

3. So we made it - but still?

As described above, we have developed a lot of tools, covering key areas for indigenous languages. Still, our tools are not for everyone, not because they are not available – all our tools are open source and free for everyone – but because the platforms and environments people are using are not open for all languages.

There are numerous examples of such hindrances, from the very low-level language codes to the highest level speech processing API's. The following are but a few of the issues we have met.

For example, our spelling checkers do not work in Chromebooks, and it is not possible to make our spellers work in Google Docs the same way as Google's own tools with red squiggles and right-click functionality. The same is true for MS Office 365. These are problems created by Google and Microsoft – we know that our spellers work, but we are not allowed to put them to use. There might not be an API for providing speller services (e.g. in Chrome OS), or Microsoft and Google have not made available systems for adding third-party spellers to their web-based Office suites.

Another example is combining diacritics in Unicode. These are often unreadable when used in indigenous languages, cf. Figure 2. The explanation is an example of colonial structures manifested in basic language technology: the Unicode consortium has established a principle that no new precomposed letters can be added to the standard, and that new base character + diacritic combinations should be handled by a dynamic composition mechanism – the diacritic is automatically placed on the base character in the optimal position.

Although this sounds like a nice way of keeping the size of Unicode within managable boundaries, in practice it does not work, and it hits indigenous and minority languages only. The problem is that the combining machinery varies by each implementor, giving inconsistent results. One can never trust that the visual appearance on one own's computer will be the same as on the reader's. And often the result is just gibberish.

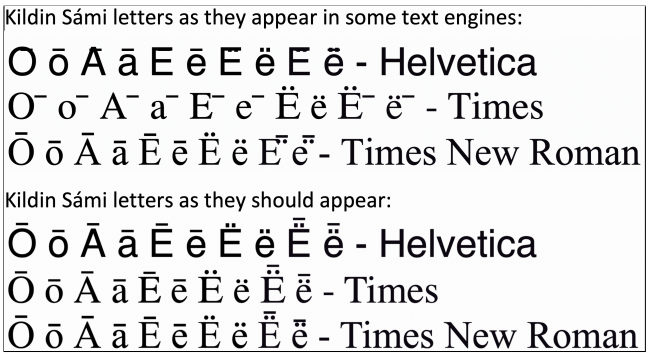


Figure 2: How dynamic diacritics often render in popular text engines (top). Correct rendering below.

Why is it only hitting minority and indigenous languages? Because all majority languages are covered by the existing precomposed letters, so none of the developers really get to see the problems in daily use. It is clearly so, since the bugs illustrated above have been consistent and persistent for at least 10 years.

While it is easy to criticise the makers of the text ren-

[4]Cf. Divvun, `http://github.com/divvun`, see also (Antonsen, 2012), (Wiechetek et al., 2019)

[5]`http://oahpa.no`

[6]For machine translation we use the Apertium formalism (`http://github.com/apertium`), cf. (Antonsen et al., 2017) for a presentation

[7]`http://gtweb.uit.no/korp`

dering engines, one could as well blame the Unicode standard for the situation. Why should new precomposed letters be banned from the standard? Space is cheap nowadays, and adding some extra letters based on existing symbols to a font should also be quite cheap. One could assume that in high-quality, broad-coverage fonts the issue would then be resolved. One could still keep the present, dynanmic composition as a fall-back for less developed fonts, and there will still be cases where dynamic compounding is an acceptable solution. Seen from the perspective of minority languages, whether the issue is solved by better font rendering or by extending the characters repertoire of Unicode is not important. The point is that it should be solved, and not be kept in a situation where neither solution works.

Dictionaries and easy access to definitions, translations and usage examples are very important to minority and indigenous language communities, much more so than for majority languages. And the ability to look up a word in text and immediately get an explanation and translation is crucial in language revitalisation, all the while these languages typically have complex inflections making a direct lookup based on the word form often challenging.

Many operating systems, both desktop and mobile, do have such functionality built-in. Unfortunately it is often either locked down, or of very limited use because there is no way to provide morphological analysis and disambiguation of the input text. So while there is plenty of support for English speakers wanting to learn or needing help in understanding German on e.g. iOS, there is no way the same service can be provided to indigenous and minority languages. It *is* possible to provide *similar* functionality via other means, but that includes extra steps for the users, steps that in practice is a blocker for actual use. Why can't there be a dictionary lookup API available to anyone and with hooks for morphological analysis, providing user services the same way as Apple's own licensed content? Why does third party content have to be treated like third class citizens?

There are a number of settings in which it is desirable to have localised software, including whole operating systems. But most software is not easily localisable, or independently localisable at all. And if it is, it is usually a non-trivial task to distribute the localisation to the users. And if you are able to overcome all these hurdles, you will eventually find that your language is not listed in the language preferences of your operating system, so that the localisation is either unavailable, or the localised software has to resort to special settings to make it available. On top of that the localisation process varies by software and operating system, and is very time consuming despite software text strings usually following a simpler syntax and a lot of repetition. The end result is that most software and operating systems are not localised beyond the dominating languages, and often they can't be localised even if one wants to.

Speech technology is one of the hotspots of language technology these days, and there are academic papers on how speech technology can help overcome the digital divide for indigenous languages by just skipping the written mode, cf. (Palkar et al., 2012). The problem is just that – even if you succeed in building that fantastic speech recognition + machine translation system – you can't make it work where the users are. That is, you can't add your own voice to the Android phones used by the language community, and chances are that the language as such is not even recognised by the operating system. And the speech technology API's are most likely closed behind an unfriendly license, or not available at all.

These are just some examples of issues meeting indigenous and minority langauge communities. The language technology groups at UiT are just two of many working to improve the situation for these language communities, but due to the issues mentioned above, *we just can't provide the tools and services our users want!*

To be clear, issues like the above are *not* restricted to the software providers mentioned, the providers are just examples. The problems are found everywhere in the software industry, including in major open-source projects.

4. Solution: A Manifesto for Open Language Technology

To solve the issues discussed in the previous section, we propose a set of simple software development principles, dubbed a *Manifesto for Open Language Technology*. The four principles are:

Open localisation: all software should be localisable independently of the producer of the software

Open interfaces: all language-related programming interfaces should be open by default

Open resources: all language resources should be open and accessible for everyone, given the permission of the language community

Accessible standards: language-related international standards should be respected, fully implemented and implementations should be regularly updated

We'll elaborate on these principles in the following sections.

4.1. Open localisation

The software belongs to its creators, but one could argue that the user interface language belongs to the language community, or rather, that any langauge community should have an independent right to localise any software they deem necessary to their language, *without asking the creator of the software*. In fact, the localisation of software and access to localisations should be made such that the software creator should not need to be involved at all.

Given that the major software platforms nowadays have their own app stores, it should be possible to add something like a *locale store* to it, a place where users can get and install localisations for any language they want. And the localisation packages should include localisations for any piece of software that has been localised into that language, be it the OS itself or any first or third party software package.

The first principle says that the language belongs to the language community, and that should also be true for localisation.

4.2. Open interfaces

Modern digital devices usually come with a lot of linguistic features, from spellers to digital assistants. Most, if not all, of these features have an API at some level. But these API's are not equal: some are open and free, such as speller engine API's, others are behind license bars, such as a lot of speech technology API's, and some are not public at all, such as the API's for adding speech assistant support for new languages.

The major software houses will never make language technology for most of the world's languages, and that is fine, they don't have to. What is *not* fine is that they don't allow the language communities to develop that technology themselves, by locking all needed API's behind bars of various kinds.

The second manifesto principle says that all API's related to language technology should be open and accessible to any language, no questions asked.

This principle does not mean that the language technology itself has to be open, it just says that if the OS vendor does not provide support for language A, it should not stop others from providing that support, and the support should be transparent for the users.

4.3. Open resources

Building linguistic resources for a language is time consuming independently of the actual technology being used. It is vital that the language resources belongs to and are in control of the language community, both for legal and ethical reasons. Too many times it has happened that language resources for an indigenous language has been owned by a private entity, blocking reuse of those resources in settings not directly benefitting that private entity.

The best way to ensure this is to always develop linguistic resources using an open license, although that must in the end be decided by the language community. The third manifesto principle says exactly that: language resources should be open independently of private entities, and the license and openness should be decided by the language community.

4.4. Accessible standards

There are a number of digital standards relating to human languages. And the standards usually take all languages of the world into consideration. But unfortunately these standards are not always implemented in full, and they are thus unreliable and unaccessible.

The most visible example is the ISO 639 series of language codes, especially the 639-3 language codes that cover in principle every language on earth. What is lacking is support from the software industry, especially the operating system developers.

All OS's do recognise these codes as language codes, but that's it. For most languages, its code is not recognised, just that it is *some* language code. This means, for example, that most languages:

- are not known to the system, and can't be a preferred locale
- show up as language *codes* if you install tools for them, not with the language *name*
- have to be hidden behind the *other* language code to be recognised on some systems
- can not be used for speech technology applications

All digital standards relating to human languages, especially the ISO 639 standard, should be treated the same way as Unicode: be fully implemented, and updated regularly both by the standard bodies and by the OS vendors. This is what the last principle is all about.

5. Conclusion

We have developed an infra and tools for a number of morphologically complex languages. We have shown that LT tools are possible for any language, irrespective of available corpus and grammatical complexity, but we have also shown that there is a lot of issues left, issues caused by lack of support and direct neglect from the major players in the software industry. Finally, we propose a short but pointed list of software development principles to address the issues we have identified. The ultimate goal is to achieve *indigenous self-determination in the digital realm.*

6. Acknowledgements

We would like to thank our collegues at Divvun and Giellatekno in Tromsø for the fruitful cooperation over the years, our international partners, a.o. The Techno Creatives, Gramtrans, Trigram, Kvensk institutt, Oqaasileriffik, and FU-lab (Syktyvkar). The work is mainly financed by The Arctic University of Norway and by the Norwegian Ministry of Local Government and Modernisation. Important steps forwards have been made possible by project support from the Research Council of Norway and by Kone Foundation of Finland, as well as by the Social Sciences and Humanities Research Council of Canada.

7. Bibliographical References

Antonsen, L., Gerstenberger, C., Kappfjell, M., Rahka, S. N., Olthuis, M.-L., Trosterud, T., and Tyers, F. M. (2017). Machine translation with North

Saami as a pivot language. In *Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa, 22–24 May 2017, Gothenburg, Sweden*, volume 29 of *NEALT Proceedings Series*, pages 123–131. Linköping University Electronic Press.

Antonsen, L. (2012). Improving feedback on L2 misspellings – an FST approach. In *Proceedings of the SLTC 2012 workshop on NLP for CALL, Lund, 25th October, 2012*, volume 80 of *Linköping Electronic Conference Proceedings*, pages 1–10. Linköpings universitet.

Arppe, A., Lachler, J., Trosterud, T., Antonsen, L., and Moshagen, S. N. (2016). Basic language resource kits for endangered languages: A case study of Plains Cree. In *Proceedings of the LREC 2016 Workshop CCURL 2016 – Towards an Alliance for Digital Language Diversity*, pages 1–8. LREC.

Beesley, K. R. and Karttunen, L. (2003). *Finite State Morphology*. Studies in Computational Linguistics. CSLI Publications, Stanford, California.

Johnson, R., Antonsen, L., and Trosterud, T. (2013). Using finite state transducers for making efficient reading comprehension dictionaries. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013); May 22–24; 2013; Oslo University; Norway*, volume 16 of *NEALT Proceedings Series*, pages 59–71. Linköping University Electronic Press.

Karlsson, F., Voutilainen, A., Heikkilä, J., and Anttila, A. (1995). *Constraint Grammar. A Language-Independent System for Parsing Unrestricted Text*. Natural Language Processing. Mouton de Gruyter, Berlin, New York.

Karlsson, F. (1990). Constraint grammar as a framework for parsing running text. In *COLING '90 Proceedings of the 13th conference on Computational linguistics*, volume 3, pages 168–173, Helsinki.

Koskenniemi, K. (1983). *Two-level Morphology. A General Computational Model for Word-forms Production and Generation*, volume 11 of *Publications of the Department of General Linguistics*. University of Helsinki.

Moshagen, S. N., Pirinen, T., and Trosterud, T. (2013). Building an open-source development infrastructure for language technology projects. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013); May 22–24; 2013; Oslo University; Norway*, number 16 in NEALT Proceedings Series, pages 343–352. Linköping University Electronic Press.

Palkar, S., Black, A., and Parlikar, A. (2012). Text-to-speech for languages without an orthography. In *Proceedings of COLING 2012: Posters*, pages 913–922, Mumbai, India, December. The COLING 2012 Organizing Committee.

Wiechetek, L., Moshagen, S., Gaup, B., and Omma, T. (2019). Many shades of grammar checking – launching a constraint grammar tool for north sámi. In Eckhard Bick et al., editors, *Proceedings of the NoDaLiDa 2019 Workshop on Constraint Grammar: Methods, Tools and Applications, Turku, Finland*, volume 33 of *NEALT Proceedings Series*, Linköping, Sweden. Linköping University Electronic Press.

Proceedings, pages 384–387
Paris, UNESCO Headquarters, 5-6 December, 2019.

Democratizing Access to Information : An Open and Inclusive Localization Model

Amel Fraisse
Univ. Lille, EA 4073 - GERiiCO
F-59000 Lille, France
amel.fraisse@univ-lille.fr

Abstract

We describe an open and inclusive localization process promoting the right of all people to use software in their mother tongue. Currently, the translation of textual resources in software is entrusted only to professional translators. This makes the localization long, expensive and intended to profitable languages. This current workflow seems impossible to apply for endangered languages for reasons of cost, and lack of professional translators. Our proposal aims at involving end users in the localization in an efficient way: while using an application, users knowing the source language (often English) could translate strings of the interface in their native language.

Keywords: Software Localization, Endangered Languages, Open and Inclusive Localization Process

Résumé

Tuvuga ubuhinga burekurira abantu bose atavangura gukoresha ubuhinga bwa none mundimi zabo kavukire. Muri kino gihe, ihindurwa ry'ibisomwa bikoreshwa murubwo buhinga bwa none rikorwa n'abahinga bo guhindura indimi babigize umwuga. Ivyo bigatuma iryo hindurwa rifata umwanya muremure, rikaba rizimvye ndetse bigatuma riba kenshi na kenshi ku ndimi zikoreshwa cane. Ico gikogwa gisa n'ikidashoboka kundimi zigeramiwe kumvo z'uburyo hamwe no kumvo zubukene bw'abahinyanyuzi b'izo ndimi. Twebwe ico dusaba n'uko abakoresha izo ndimi bohabwa urihara rukomeye muri iryo hindurwa ry'indimi, na cane cane ko ababukoresha bazi izo ndimi busanzwe buteguwemwo (kenshi na kenshi icongereza) boshobora kugerageza kubuhindura bakabushira mu ndimi z'iwabo kavukire.

1. Introduction

Currently, the translation of technical documents as well as user interface strings is entrusted only to professional translators. In practice, software publishers send original versions of the files to be localized to several professional translators. Each translator translates and sends the translated versions to the publishers. But, it seems impossible to continue in this way for most endangered languages, for reasons of cost, and quite often scarcity or even lack of professional translators (costs increase while quality and market size decrease). On the other hand, free software such as that produced by Mozilla [1] is translated by volunteer co-developers into many (more than 90) languages, in some cases more languages than commercial software. The software localization is based on the contribution of volunteers (Tong, 1987; Vo-Trung, 2004; Lafourcade, 1991). Another situation (different from the translation of technical documentation) is that of occasional volunteer translators, who contribute without an organic connection to the project. Hence, it is possible to obtain high quality translations of documents that may be over a hundred pages long (such articles of the Wikipedia encyclopedia, texts of Amnesty International and Pax Humana). Another problem of the classical localization process is that strings of the interface are often translated out of context. Hence, the choosing the appropriate translation is not always possible due to lack of context, and in such cases even a professional translator cannot produce a perfect translation. As proposed in (Boitet, 2001; Fraisse, 2010), one solution to this problem is to involve end users with a knowledge English and who, during the use of software products, translate or improve some translations proposed by machine translation (MT) systems or translation memory (TM) systems.

1.1. Current Situation

1.1.1. Crowdsourcing work force

Many online localization communities are formed and managed by volunteer localisers, software engineers, end users and in general people sharing the same motivations and aims. As a result, many projects aimed at the translation and localization of open source software and associated documents have appeared. Two quite interesting projects of this type are: the *Mozilla* localization project [2] and the Ubuntu LoCo project [3]. The *Mozilla* software set (the web browser and the email client) is available in more than 280 languages including under-resourced ones. The *Mozilla* localization model is a continuous process because each new version has new documentation and a new interface that must be translated.The Ubuntu local community team (LoCo) project is another example of a successful localization project performed by volunteer contributors. It involves 204 official local communities. Some of these communities are linked by country (Indonesia or Germany for example), others are linked by language (Catalan or Kurdish, for example), others by a geographic location (Austin, Texas USA, or Bangalore in India). Local communities involved in the localization process decide what

[1] https://l10n.mozilla.org

[2] https://www-archive.mozilla.org/projects/l10n/

[3] https://loco.ubuntu.com

needs to be localized, how and when. Translation management is entirely online through a web interface called *Launchpad Translations*. To participate in the localization of Ubuntu, volunteer contributors must be registered on the website to identify which projects are being translated, and in the context of these projects, what specific strings have to be translated. Once registered on the website the contributor can provide a translation in his native language. In fact, very often, different contributors propose several translations for the same text. But only translators ranked as *experienced* (the official members of core translators) have the right to *validate* the translations, which will then be included in *Ubuntu*.

1.1.2. Increasing demand and need for localization

Currently, in the case of commercial software, the localization decision is driven exclusively by the economic imperative. For reasons of cost, digital publishers are obliged to localize their product only to viable markets. On the other hand, the demand for localization is growing all over the world at a very fast pace, demographics and business globalization has forced the rest of world to adopt new technologies such as the Internet, wireless networks, global communications and computers. Consequently, the demand for globalization and localization is increasing. How localization will be performed and in which language is another issue. According to Sapient Globalization Report there are over 6,700 living languages in the world; the fifteen most popular languages are spoken by 49.5% of the world's population, while the other 51.5% of the world's population speak 6,600 languages. Yet, only about 6% of the world's population speak English.

2. The alternative : An open and inclusive localization model

2.1. Basic Principles

The proposed localization model is based on two basic principles:

2.1.1. Inclusive: involving volunteer translators and end users in the localization process

As we have said above, localization seems impossible for most endangered languages for reasons of cost, and quite often a scarcity or even lack of professional translators. Our solution aims at involving non-professional translators such as volunteer localisers and especially end users. These groups have the capacity to participate effectively, since they have a better knowledge of the target language (generally their native language) and of the context of use of the software. In order to motivate this type of translators and to give them a better knowledge about the use context of User Interface (UI) strings, localization should be carried out while using the software.

2.1.2. Open: from close, discontinuous, coordinated and out-of-context localization to open, continuous, uncoordinated and in context localization

Our solution aims to move from a close, discontinuous, coordinated and out-of-context localization model to open, continuous, uncoordinated and in context localization model. The basic concept consists of renouncing the idea of perfect translation and publishing rough translations with a variable quality, which will be improved incrementally during the use of the software. Therefore, the translation process will be on going and improve continuously. This solves the problem of time since users do not have to wait for the final localized version in their language. They can download, at any time, a partially localized or non-localized version of the software. Similarly, the software publisher may first publish a partially localized version that will be progressively localized through use, leading, eventually, to a complete localized version. So, the new process permits the incremental augmentation of both quality and quantity. The same principle already exists and is used by many translation communities. The best known is the Wikipedia Community, when content is added and translated continuously by contributors. So, the idea is to extend this principle to the localization filed.

2.2. Global approach

In (Fraisse, 2010), we proposed an alternative paradigm that will permit end users, volunteer translators, etc. to take part in the localization process in an efficient and dynamic way: while using the software, the end users who know the source language of the software (often but not always English) can translate or improve the previously existing translations. It is therefore an open and inclusive localization model. However, the publisher may ask professional translators and reviewers to translate the crucial parts of the software. The proposed localization model can be performed individually or collaboratively. In fact, the user has the choice to localize his/her software locally without any exchange with other users or to localize collaboratively.

2.2.1. Localizing locally

In a previous work (Fraisse et al., 2009), we have proposed a local use of our alternative localization model. That is, the model allow user to translate any string of the user interface locally. The scenario that we have envisioned is that: The user right-clicks on any string of the interface, which brings up a context menu Figure 1. That allows the user to localize the string. This is achieved by: (i) Entering a new translation, or choosing one of the proposed translations, (ii) Clicking on the *Localize* button, (iii) The interface updating in real time.

As a result, the user does not have to be connected to any resource or platform. All new translations proposed by the user are saved to a local resource file. On their next connection to the collaborative platform (we present and describe the role of the collaborative platform in the next subsection), the resource file will be analyzed and synchronized. In this way the user can submit his/her translations and get suggestions of translation proposed by other users.

2.2.2. Localizing collaboratively

End users can also localize online in order to exchange and access linguistic resources such as glossaries, dictionaries, translation memories, machine translation systems, etc. that are available on a collaborative platform. We have envisioned two possible collaboration scenarios: the first

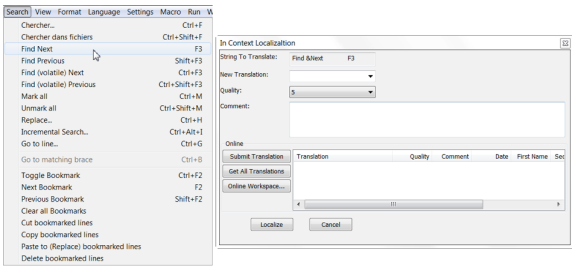


Figure 1: In context localization of the string *Find Next* of the software *Notepad++*

consists of interacting with a collaborative platform during localization; the second scenario requires localization to take place directly on the collaborative platform.

First Scenario: localization through interaction with a collaborative platform

The user right-clicks on any string of the UI to allow the string to be edited (Figure 2) : (i) The collaborative platform displays all translations proposed by other users, (ii) The user enters a new translation, or chooses one of the proposed translations, (iii) The translation is submitted to the localisers site, (iv) The user clicks on the *Localize* button, (v) The interface is updated in real time.

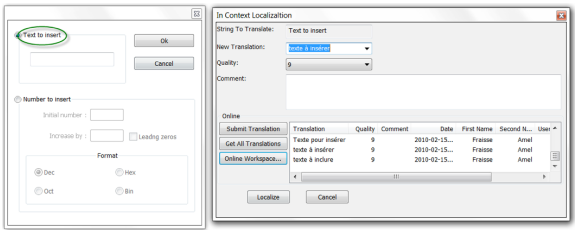


Figure 2: Collaborative and in context localization of the string *Text to insert* through interaction with a collaborative platform

Second Scenario: localizing directly on the collaborative platform

The user right-clicks on any string of the interface. This allows editing of the string directly on the collaborative platform Figure (3): (i) the user is redirected to the collaborative platform containing the string that has been chosen for translation (more details about the interface and the functionalities of the collaborative platform are described in (Huynh et al., 2008)). (ii) the user enters a new translation, or chooses one of the proposed translations. (iii) he/she returns to the original application. (iv) the interface is updated in real time.

2.3. Technical solution for an open and inclusive localization model

To enable open and inclusive localization for existing software, it is necessary to perform an internal intervention on the source code. To be as generic as possible and modify

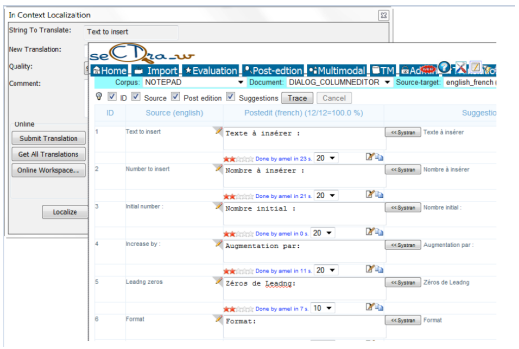


Figure 3: Collaborative and in context localization of the string *Text to insert* through interaction with the *Sectra_w* collaborative platform ((Huynh et al., 2008)).

the source code as little as possible, our modifications are only carried out on the base classes that generate all graphical user interfaces (GUIs) of the application. These modifications consist of adding new behavior adapted to the in context localization process to strings of the UI: through a simple right-click on a string in the interface, the user can choose from a list of possible translations and can add a new one, which is then updated in real time. We have implemented an open localization module pluggable into the architecture of any new or existing software developed using an object oriented programming language. The application interacts with the open localization module during the editing of user interface strings and during the update of the user interfaces. More details about the implementation of this module can be founded in our previous works (Fraisse et al., 2009, Fraisse, 2010).

3. Experiments and Results

The proposed localization model was experimented to localize the free software *Notepad++* initially localized into only 55 languages. We have decided to localize it into Vietnamese (not included in the 55 languages) and so, have created the source and the target corpus in order to import them into the *Sectra_w* collaborative platform (Huynh et al., 2008). The source corpus is based on the resource files of *Notepad++*. In the case of *Notepad++*, all strings of the user interface are stored in a single resource file named *NativaLang.xml*. We extracted and stored the UI strings in a textual file of the source corpus. To build an initial target corpus, we used *Google-Translate* to translate the UI source strings. *Notepad++* has 600 user interface strings. Three Vietnamese native speakers and users of *Notepad++* were selected to participate in the experiment. An open localizable version of *Notepad++* was installed on each of their computers and an account for each user was created on the *SECTra_w* collaborative platform (Huynh et al., 2008). For each participant the progress of his work was observed and they were asked to report the amount of time that they spent on the localization of the application. The results of this experiment are shown in Table 1.

In total the three users spent about eight hours to translate the 600 strings of the UI. We asked a professional trans-

lator about the time required to translate 600 strings form English to Vietnamese and the answer was about six hours. To assess quality of the translations, we sent the *NativeLang* file (containing the all source strings of the UI) to a professional translator and then we have asked the three user-translators to compare their translations with those performed by the professional translator. The conclusion of this comparison allowed us to deduce that the professional translator had translated approximately 36% of UI strings poorly. We attribute this to the fact that the UI strings were translated out of context.

	Day 1		Day 2	
	Time	Nb. str.	Time	Nb. str.
user 1	40mn	80	140mn	120
user 2	60mn	115	160mn	126
user 3	60mn	100	30mn	59
Total	160mn	295	330mn	305

Table 1: Duration and number of translated strings by user

4. Conclusion

We have proposed an open and inclusive localization model for new and most existing software. This new process offers a new solution to allow software localization into endangered languages and thus promoting the right of all people to use software in their mother tongue. The proposed model is an alternative to the current localization process that seems impossible to apply for most endangered languages for reasons of cost, and quite often a scarcity or even lack of professional translators. It includes volunteer localisers and specially end users in the localization process in an efficient and dynamic way: while using a software (in context), users, knowing the current language of the user interface strings, can right-click on strings to translate or improve translations. This model was experimented to localize into Vietnamese the *Notepad++* software. The localization experiment was successfully accomplished. In total the 600 strings of the user interface were localized by 3 volunteer Vietnamese native speakers.

5. Bibliographical References

Boitet, C. (2001). Four technical and organizational keys for handling more languages and improving quality (on demand). In *Proceedings of MTS2001, IAMT*, Santiago de Compostela, September.

Fraisse, A., Boitet, C., Blanchon, H., and Bellynck, V. (2009). A solution for in context and collaborative localization of most commercial and free software. In *Proceedings of LTC 2009 the 4th Language and Technology Conference, vol. 1/1*, pages 536–540, Poznań, Poland, November.

Fraisse, A. (2010). *Localisation interne et en contexte des logiciels commerciaux et libres*. Thèse de doctorat en informatique, Université de Grenoble.

Huynh, C., Boitet, C., and Blanchon, H. (2008). Sectra_w.1: an online collaborative system for evaluating, post- editing and presenting mt translation corpora. In *In proceedings of 6th Language Resources and Evaluation Conference*, Marrakech, Morocco, May.

Lafourcade, M. (1991). Odile-2, un outil pour traducteurs occasionnels sur macintosh. In *Presses de l'université de Québec*, Université de Montréal.

Tong, L. (1987). The engineering of a translator workstation. *Computers and Translation*, pages 263–273.

Vo-Trung, H. (2004). *Méthodes et outils pour utilisateurs, développeurs et traducteurs de logiciels en contexte multilingue*. Thèse de doctorat en informatique, Institut National Polytechnique de Grenoble.

Proceedings, pages 388–392
Paris, UNESCO Headquarters, 5-6 December, 2019.

The Case of Polish on its Way to Become a Well-Resourced-Language

Zygmunt Vetulani[1], Grażyna Vetulani[2]
Adam Mickiewicz University in Poznań[1,2]
ul. Wieniawskiego 1, 61-712 Poznań, Poland,
{vetulani, gravet}@amu.edu.pl

Abstract

We intend to illustrate – on our own research – the effort one needs to invest in order to obtain a real size application of deep text understanding for a language being categorized as *less-resourced.* We present some language resources and tools we had to develop for processing text in Polish, a language which was still in this category in the 1990s. The resources were used to implement a full scale prototype of the rule-based system POLINT-112-SMS for improving information workflow for emergency management purposes.

Keywords: less-resourced-language, Language Technology, language resources, dictionaries, lexicon grammars, tools, applications with NL competence.

1. Introduction

Polish, a West Slavic language spoken by about 45 million native speakers, has written literature since XIII century and in XXth century was among the best described languages. Still it was classified for a long time as a language with a very poor technological infrastructure. Using today's terminology, Polish was considered as a "less-resources-language" at least until the late 1990s. Negative effects of this scarcity time may still be observed. The EU initiatives of of awareness actions followed by appropriate funding measures appeared to be an important milestone in the development of human language technologies (HLT) in Central Europe. Successive incorporation to the international (and multilingual) research was effect of new bridges emerging between the technologically advanced research communities and newcomers often contributing with new original ideas. International grants bringing together partners from various language communities and representing different linguistic traditions were often oriented to creation of basic language resources and tools requested by quickly growing multilingual sector of language industries, also in Poland.

The general condition of language technologies for Polish is now good enough to no classify Polish as "less-resourced". Having said that, there are still many gaps to be filled concerning language resources and tools.

The conference-forum LT4ALL 2019 organized under the auspices of UNESCO that has just ended in Paris was today's equivalent of the EU actions taken a quarter of a century ago. It was addressed primarily to the representatives – researchers and administrative staff – of clearly "less-resourced-languages", but also to other LT community members, in order to sensitize all of them to new threats of a "world-at-various technological speeds".

In this paper we want to share our experience in the ambitious task to acquire the capacity to develop LT based technologies in the situation of initial scarcity of digital language resources and tools for Polish.

2. Beginnings

Intensive development of Polish literature since Renaissance and Baroque was followed by high quality linguistic descriptions of the language, so that in 20th century it was among the best described languages. Rapid progress of computer technologies attracted the attention of researchers on possibility of machine translation already in late 1940s. In Poland, among the first attempts to use computers in processing Polish text and speech worthy a note are works conducted-in 1970s in Warsaw and Poznań (L. Bolc, S. Szpakowicz for text processing, W. Jassem, M. Steffen-Batóg for speech generation and analysis). In late 1970s and early 1980s the first attempts to implement toy systems understanding Polish were done (independently) by W. Lubaszewski, St. Szpakowicz, Z. Vetulani. The common feature of these first works was inaccessibility of real-size electronic language resources and NLP dedicated tools for Polish. Still, these initial works (see e.g. Z. Vetulani, 1984 and 1988) appeared stimulating a decade later when we started working on real size application prototypes of systems involving text understanding (Vetulani, 1991). These initial works highlighted necessity of real scale language resources.

3. Basic resources for text processing

On the ground of our trials to design non-trivial question-answering systems with deep understanding (rule based) (Vetulani, 1984, 1988) we realized that the most urgent necessity was to dispose of electronic dictionaries, computer processible grammars of Polish and corpora (both application specific, but also corpora for general language). Several projects contributed to fill (partially) the lack of the above mentioned resources and tools.

3.1 Dictionary project POLEX (1994-1996)

Polish is a language with a complex inflection system and has relatively free word order. Therefore simple adaptation of processing algorithms efficient for English or French appeared hard to apply, as basic information concerning the function of a word in the sentence is typically being encoded in the word form, independently of its linear position in the sentence.

Good grammatical description of Polish existed until recently only in form of traditional dictionaries and grammars addressed to the human users and were of low usefulness for automatic processing because of lack of precision. Huge amount of work invested by grammarians

until 1990s did not lead to a standard of description of Polish words that would eliminate the necessity of individual linguistic competence of users to interpret dictionaries. We proposed unambiguous inflectional description apt to eliminate linguistic competence of dictionary users and therefore be appropriate for machine text processors. The solution we proposed is an outcome of the POLEX Polish Lexicon Project (1994-1996)[1].

The POLEX Polish Lexicon is a morphological dictionary which includes the core Polish vocabulary of general interest of the traditional paper dictionary (Szymczak, 1981). It is based on a precise machine-interpretable formalism (coding system), the same for all categories (classes of speech) (Vetulani et al., 1998).

The dictionary entries are of the following form:

BASIC_FORM+LIST_OF_STEMS+PARADIGMATIC_CODE+DISTRIBUTION_OF_STEMS

The paradigmatic inflection code contains full paradigmatical information, i.e. the way how to associate endings to stems to obtain a required form of the word. The distribution associates stems to the paradigmatic positions.

The first public release of the resource contained over: 42,000 nouns, 12,000 verbs, 15,000 adjectives, 25,000 participles, and about 200 pronouns[2].

3.2 First steps towards Lexicon grammars for Polish

In the early 1970s Maurice Gross (LADL, Paris 7) the concept of a grammatical lexicon based on the idea of storing words together with possibly all relevant syntactic and semantic information (Gross, 1975). This idea, inspired by Z. S. Harris' transformation theory, was developed first for French, then for other languages. Consequently, predicative words were studied from the point of view of their aptitude to form elementary sentences. Gross introduced the term lexicon-grammar (fr. lexique-grammaire) to mean the method to describe the meaning of predicative words by providing description of how these words form simple sentences. What distinguishes lexicon grammars from traditional grammatical descriptions of a language is that lexicon grammar entries contain possibly full grammatical description[3] of well distinguished senses of words. This property makes that lexical-grammars are adequate for application in language processing systems.

The EUREKA project GENELEX (1990-1994)[4] was an initiative to define a generic model for lexicons; to design and develop software tools for lexicon management (Marie-Hélène ANTONI-LAY et al., 1994) based on the ideas of lexicon-grammar. Anoni-Lay gives two reasons to build large size lexicons. "The first reason is that Natural Language applications keep on moving from research environments to the real world of practical applications. Since real world applications invariably require larger linguistic coverage, the number of entries in electronic dictionaries inevitably increases. The second reason lies in the tendency to insert an increasing amount of linguistic information into a lexicon. (…) In the eighties, new attempts were made with an emphasis on grammars, but an engineering problem arose: how to manage a huge set of more or less interdependent rules. The recent tendency is to organise the rules independently, to call them syntactic/semantic properties, and to store this information in the lexicon. A great part of the grammatical knowledge is put in the lexicon (…). This leads to systems with fewer rules and more complex lexicons." (ibid.).

Two COPERNICUS projects CEGLEX – COPERNICUS 1032 (1995-1996) and GRAMLEX – COPERNICUS 621 (1995-1998) were executed under the EC funded 4th Framework Program which accepted participation of Central European Countries. One of the goals of these projects was to test the potential of the extension of novel LT solutions to languages that were going to be considered as official in future EU members after 2004.

3.2.1 PECO-COPERNICUS project CEGLEX (1995-1996)

The main goal of the CEGLEX consortium (Vetulani et al. 1994) was to test the GENELEX proposal of a generic model for re-usable lexicons – first implemented for a number of West-European languages, among other French, English, German, Italian – for three more languages spoken in Central Europe: Czech, Hungarian and Polish.

The CEGLEX/GENELEX model claims to be:

- theory-welcoming,
- complete, i.e. to cover all relevant phenomena on three classical layers: morphological, syntactical, and semantic,
- easily transportable.

The three layers of the CEGLEX/GENELEX model were confronted with the data of the considered languages with generally positive results, especially for Czech and Polish. For Polish this confrontation consisted in the adaptation of the model to the Polish data. On this occasion some modifications were proposed, in particular concerning the representation of the inflection phenomena. It is worth noticing that the POLISH CEGLEX module went further than GENELEX as we considered also semantic layer which was only marginally addressed in GENELEX. The outcome of CEGLEX was the first successful attempt to test on representative linguistic data feasibility of machine readable lexicon-grammar covering all three layers.

3.2.2 PECO-COPERNICUS project GLAMLEX (1995-1998)

The aim of the COPERNICUS Project 621 GRAMLEX was to facilitate the initiation, coordination and standardization of the construction of morphological dictionary packages for the following European languages: French, Hungarian, Italian and Polish, including detailed formal description of the morphology of the languages. The

[1] Resarch project „POLEX - Polska Leksykalna Baza Danych No KBN8S50301007" realised by Z. Vetulani, B. Walczak, T. Obrębski, G. Vetulani and other team members (1994-1996).

[2] The resource is distributed through ELRA. ISLRN: 147-211-031-223-4; ID: ELRA-L0047

[3] On both syntaxic and semantic levels.

[4] GENELEX was continued by LE-PAROLE (1996-1998), LE-SIMPLE (1998-2000) and GRAAL (1992-1996) projects.

intention of the GRAMLEX tasks as to Polish was to contribute to the improvement of the situation concerning language engineering tools and resources for Polish. Among the main achievements was a corpus-based SGML-encoded (in format GRAMCODE) morphological dictionary of Polish (over 22.500 entries) and related tools and applications (lemmatizer, inflected form generator, concordance generator and other5). The project GRAMLEX was closely connected with two projects mentioned above POLEX and CEGLEX.

3.3 Further steps towards Lexicon grammars for Polish. PolNet 3.0 as Lexicon-Grammar

The IT applications with language competence we were able to develop until 1990 were all of the category of toy-systems. This was, first of all, because of scarcity of real-size digital, easily machine processable electronic resources. This problem was addressed in the R&D grant "Text Processing Technologies for Homeland Security Purposes" that we realized during 2006-2010[6]. Within this grant we created a prototype of POLINT-112-SMS system designed to assist the monitoring process of mass events and to enhance real-time identification of processes in the crowd of fans in order to discover potentially dangerous situations with a high degeneration risk (early prevention).

The POLINT-112-SMS project required a robust natural language competence to understand and process SMS messages exchanged between the security staff agents in uncontrolled natural language (cf. Vetulani and Osiński, 2017). The understanding module of the system is rule-based because of necessity to obtain very precise representation of utterance content which is crucial in processing neuralgic information. Messages were supposed to be written in standard, correct and unconstrained Polish.

3.3.1 PolNet - Polish Wordnet as lexical ontology (since 2006)

Within the project POLINT-112-SMS (Vetulani et al., 2010) we made extensive use of ontology to represent meaning of text messages. Absence on the market of lexical ontologies reflecting conceptualization typical of Polish speakers inspired us to develop PolNet Polish Wordnet – a lexical database of the type of Princeton WordNet[7]. We built it from scratch for Polish following the so called "merge model" methodology[8] PolNet design started in 2006 and its progress continues. The resource development procedure was based on the exploration of good traditional dictionaries of Polish and the use of available language corpora (e.g. IPI PAN Corpus; cf. Przepiórkowski, 2004). Development of PolNet was organized in an incremental way, starting with general and frequently used vocabulary[9].

We decided to selected the most widely used words found in a reference corpus of Polish language (ibid.) with however one important exception made for methodological reasons. Even though we wanted the core of PolNet to be a resource of general interest, we also assumed its possibly early validation in the real-size applications.

By 2008, the initial PolNet version based on noun synsets related by hyponymy/hyperonymy relations was already reach enough to serve as core lexical ontology for real-size applications. However, to develop a POLINT-112-SMS system prototype, an extension of the core set of nouns with domain terminology was necessary. Further extension with verbs and collocations transformed PolNet in a lexicon-grammar facilitating implementation of parsers.

3.3.2 From PolNet to Lexicon-Grammar for Polish. PolNet 1.0

Already in early 1980s information typically contained in lexicon-grammar entries for predicative words, simple or compound, was considered useful for parsing and generating natural language sentences. Lexical entries used in the PROLOG code of the demonstration system ORBIS[10] (Colmerauer and Kittredge, 1982), were in fact lexicon-grammar units describing syntactic and sematic valency of words. The syntactic/semantic valency was used as constraints permitting avoiding producing incorrect sentences by a generation algorithm, to avoid accepting incorrect sentences by a parser, and to build error-correcting software (ibid, see also Vetulani, 1988). Qualitative evolution of PolNet, initially conceived as lexical ontology, towards a lexicon-grammar of Polish took place at the passage from the release of PolNet 0.1 (2009) to the version PolNet 3.0. (2014). The pragmatic reason to substantially enrich PolNet was the need of an efficient parsing engine to support the understanding module.

In addition to noun synsets that make of PolNet 0.1 (2009) a lexical ontology, we decided to enrich PolNet with verb synsets containing syntactic/semantic information in form of valency structure. The valency structure of a predicative word provides the morpho-syntactic and semantic constraints on the acceptable fillers of the argument positions opened by this word (like *case, number, gender, preposition, register* etc. for morpho-syntactic constraints and semantic roles like *agent, patient, beneficient,* etc. for semantic ones). In (Vetulani, Z. & Vetulani, G., 2014b) we

[5] These tools and applications were:
1) a lemmatizer/tagger (LEXAN) (Vetulani et al., 1997;1998),
2) a generator of inflected forms for simple and compound lexemes (Vetulani et al. 1998),
3) a syntactic concordance generator (SCON) (Vetulani et al. 1998),
4) a tool for extraction of compound terms and terminology from texts (EXTRACT) (Vetulani et al. 1998),
5) an application for structure analysis of dictionary entries (VERBAN) (Vetulani et al. 1998),
6) an application for acquisition of the lexicon from dictionary definitions (NOUNAN) (Vetulani et al. 1998),
7) an application for interactive analysis of dictionary definitions (NOUNDAN) (Vetulani et al. 1998).

[6] Grant of Polish Ministry of Science and Higher Education (MNiSzW) Nr R0002802.

[7] In the Princeton WordNet (and similar systems) the basic entities are synsets, i.e. classes of synonyms related by some relations of which the most important are hyponymy and hyperonymy.

[8] Princeton WordNet (Miller et al., 1990) was used as a formal ontology to implement systems with language understanding functionality. In order to respect specific Polish conceptualization of world, we decided to build PolNet from scratch rather than merely translate Princeton WordNet into Polish.

[9] See (Vetulani et al., 2007) for PolNet development algorithm.

[10] ORBIS, an interface to a database on planets, was entierly implemented in PROLOG to show the qualities of declarative programming paradigm.

presented the idea of a verb synset as follows: "In opposition to nouns, where the focus is on the relations between concepts (represented by synsets), and in particular on hiperonymy/hyponymy relations, for verbs the main interest is in relating verbal synsets (representing predicative concepts) to noun synsets (representing general concepts) in order to show what connectivity constraints corresponding to the particular argument positions are. This approach imposes granularity restrictions on verbal synsets and more exactly on the synonymy relation. Synonymous will be only such verb+meaning pairs in which the same *semantic roles* take the same concepts as value (this is necessary but not sufficient). In particular, the valency structure of a verb is one of formal indices of meaning (members of a sysnset share the valency structure)." (Vetulani, Z., Vetulani, G. 2015).

Verb synsets appeared already in the first public release of PolNet in 2011 (PolNet 1.0) (Vetulani et al. 2016). This opened a new generation of PolNet systems that we call now "PolNet – Polish Lexicon-Grammar systems". The expansion of PolNet to Lexicon Grammar of Polish was based on the results of theoretical research on predicative verbs assembled in Dictionary of Polish Verbs (Polański, 1980-1992) where linguistic descriptions were provided for 7000 Polish predicative verbs.

The valency information permitted us to make a smart use of PolNet enriched with lexicon-grammar features when implementing the POLINT-112-SMS system. In addition to using PolNet as lexical ontology in the World Knowledge and Situation Analysis Modules, we made use of the valency information to enhance efficiency of the parser being a part of the Text Understanding Module. In this module, syntactic/semantic valency information stored in lexicon-grammar rules was used to control parsing execution by heuristics[11] in order to speed-up parsing due to additional information gathered at the pre-analysis stage. The effect of substantially reducing the processing time was due to the reduction of search space.

3.3.3 Collocations in PolNet 2.0 - PolNet 3.0.

Usefulness of lexicon-grammar approach confirmed through successful implementation of the POLINT-112-SMS system's prototype motivated us to successive extension of PolNet towards a full lexicon-grammar. The versions PolNet 2.0 and PolNet 3.0 are important milestones in this process.

The passage from PolNet 1.0 to PolNet 2.0 was marked by inclusion of an important set of verb-noun collocations from the "Syntactic dictionary of verb-noun collocations in Polish" (Vetulani, G. 2000 and 2012)[12] or directly from corpora (Vetulani et al. 2010)[13]. Adding verb-noun collocations to PolNet appeared a non-trivial task because of specific morpho-syntactic phenomena related to collocations as for example syntactic synonymy[14] (Vetulani et al. 2016), as well as the problem of (optimal) granularity of verbal synsets. In (Vetulani, Z., Vetulani, G.,2014b) we noticed: "The challenging issue of verb synsets granularity is closely connected with synonymy which is fundamental for the concept of wordnet. Let us notice the fact (…) that while there is consensus through the wordnet community concerning the principle that *synonymy is the basis of organization of the (wordnet) database in synsets*, (i.e. synonyms should belong to the same synsets), there is no consensus among linguists on the concept of synonymy. Miller and Fellbaum (in Vossen et al., 1998) postulate a very weak understanding of this concepts (based on *invariability test with respect to just one linguistic context*) often leading to very large synsets."

The version 3.0 of PolNet was meticulously cleaned and extended[15] with respect to the version 2.0. It has been user-tested as a resource for modeling semantic similarity between words (Kubis, 2015).

4. Conclusions

This article is a case study to illustrate the challenges on way to achieve ambitious technological goals and to give an idea of the effort to be invested in the situation of initial scarcity of language resources and tools – typical of less-resourced-languages. We presented our long term works resulting with implementation of a sophisticated ICT system and development of significant language technology workbench. To achieve success it was necessary to first collect or produce basic instruments. Therefore the urgent need was to produce ready-to-use resources like processable text and speech corpora, electronic dictionaries, computer-readable grammars.

By no means we pretend to claim that the solutions we present here are the only good and sufficient measures to reach the goal which is to obtain for Polish the status of a well-resourced-language. In fact, the totality of good effects obtained on this way were due to the individual or collective effort of the LT community in Poland.[16]

[11] A well-constructed heuristic permits – on the basis of morphological and valency information combined with the switch technique Vetulani (1994) – to to reduce the complexity of parsing down to linear in an important number of cases.

[12] This dictionary is described in two monographs. The first one (2000) describes the initial phase of works on a dictionary of verb-noun collocations together their usage in sentences as predicates (2862 predicative nouns). This work was done manually. Extension of resource to 14.600 collocations was described in the second book (2012) reporting further, computer-assisted work. A part of this resource was integrated with PolNet.

[13] In (Vetulani et al. 2010) we described a computer-assisted algorithm to extract collocations directly from text corpora. The algorithm requires involvement of qualified lexicographers.

[14] In Polish we observe the phenomenon of syntactic synonymy (Jędrzejko 1993) where for some predicative verbs their morpho-syntactic structure is different from the morpho-syntactic structure of their semantic synonyms in form of verb-noun collocation (e.g. for the direct complement). Therefore to be consistent with our methodological assumption, we will range these synonymous forms in distinct synsets interconnected by a special semantic similarity relation.

[15] From14.400 in PolNet 2.0 to 17.564 in PolNet 3.0.

[16] We applied some outcomes of this effort., as e.g. the IPI PAN Corpus of Polish teksts (Przepiórkowski 2004), further extended to a National Corpus of Polish Language. Other crucial resources as wordnets were still non-existing or not available at the time we needed them. For example another wordnet for Polish – Słowosieć (also known as plWordNet; see Piasecki et al. 2009) was developed independently of PolNet at about the same time and according different methodological bases.

A particular attention should be attracted to make the developed resources and tools sustainable, reusable, open for further development, and – last but not least – easy to maintain. All these require an additional effort.

Three kinds of circumstances may considerably speed-up the transition of the status of the language from "less-resourced" to "well-resourced":

- for language: to have solid traditional linguistic description,
- man-power: dispose of well-formed staff and students: linguists and computer engineers,
- technology: being eligible for receiving international assistance, participate in international development programs and projects, possibly as full partners.

5. References

Colmerauer A. and Kittredge R. (1982): ORBIS. In J. Horecký (Ed.) *Proceedings of the 9th COLING Conference*.

Gross, M. (1975): *Méthodes en syntaxe,* Paris: Hermann.

Jędrzejko, E. (1993): *Nominalizacje w systemie i w tekstach współczesnej polszczyzny,* UŚ Katowice.

Kubis, M. (2015): A semantic similarity measurement tool for WordNet-like databases. In Z. Vetulani, J. Mariani (Eds.), *Proceedings of the 7th Language and Technology Conference*, Poznań, Poland, 27-29 November 2015. FUAM, Poznań, 150 – 154.

Miller, G. A, Beckwith, R., Fellbaum, C.D., Gross, D., Miller, K. (1990): WordNet: An online lexical database. Int. J. Lexicograph. 3, 4, 235–244.

Piasecki M., Szpakowicz S., Broda B. (2009) *A Wordnet from the Ground Up*, Oficyna Wydawnicza Politechniki Wrocławskiej,Wrocław.

Polański, K. (Ed.) (1980-1992): *Słownik syntaktyczno-generatywny czasowników polskich,* vol. I-IV, Ossolineum, Wrocław, 1980-1990, vol. V, Instytut Języka Polskiego PAN, Kraków, 1992.

Przepiórkowski, A. (2004): *Korpus IPI PAN. Wersja wstępna* (The IPI PAN Corpus: Preliminary version). IPI PAN, Warszawa.

Vetulani, G. (2000): *Rzeczowniki predykatywne języka polskiego. W kierunku syntaktycznego słownika rzeczowników predykatywnych na tle porównawczym*, Adam Mickiewicz University Press: Poznań.

Vetulani, G. (2012): *Kolokacje werbo-nominalne jako samodzielne jednostki języka. Syntaktyczny słownik kolokacji werbo-nominalnych języka polskiego na potrzeby zastosowań informatycznych. Część I,* Adam Mickiewicz University Press: Poznań.

Vetulani, G. (2017): Próby formalizacji zdań opartych na predykatach rzeczownikowych języka polskiego, *Linguistica Copernicana*, 14/2017, Wydawnictwo Naukowe UMK, Toruń, 127-143.

Vetulani, Z. (1988): PROLOG Implementation of an Access in Polish to a Data Base. In *Studia z automatyki*, XII, PWN, 5-23.

Vetulani, Z. (1991): Lexical preanalysis in a DCG parser of POLISH. In E. Klein et al. (Eds.), *Betriebslinguistik und Linguistikbetrieb. Akten des 24 Linguistischen Kolloquiums, Bremen 1989*, (Linguistisches Arbeiten 260/261), Max Niemeyer Verlag, Tübingen, 389 - 395.

Vetulani, Z. (1994): SWITCHes for making Prolog more Dynamic Programming Language, Logic Programming, *The Newsletter of the Association for Logic Programming*, vol 7/1, February 1994, page10.

Vetulani, Z., Martinek, J., Vetulani, G. (1995): The CEGLEX dictionary model for Polish. In R. Bazylewicz, O. Kossak (Eds.), *Proc. of the 4th and 5th International Conferences UKRSOFT (Lviv, 1994, 1995)*, SP «BaK», Lviv, 1995, 144 - 150.

Vetulani, Z., Martinek, J., Obrębski, T., Vetulani, G. (1997): Lexical Resources and Tools for Tagging Polish Texts within GRAMLEX. In *Investigationes Linguisticae, XXI:2, 1997*, 401-416.

Vetulani, Z., Walkowska, J., Obrębski, T., Konieczka, P., Rzepecki P., Marciniak, J. (2007): PolNet - Polish WordNet project algorithm. In Z. Vetulani (ed.) *Proc. of the 3rd Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, October 5-7, 2007, Poznań, Poland*, Wyd. Poznańskie, Poznań, 172-176.

Vetulani, Z., Marcinak, Obrębski, T., Vetulani, G., Dąbrowski, A., Kubis, M., Osiński, J., Walkowska, J., Kubacki, P., Witalewski, K. (2010): *Zasoby językowe i technologie przetwarzania tekstu. POLINT-112-SMS jako przykład aplikacji z zakresu bezpieczeństwa publicznego (Language resources and text processing technologies. POLINT-112-SMS as example of homeland security oriented application)*, Adam Mickiewicz University Press: Poznań.

Vetulani, Z. (2012): Language Resources in a Public Security Application with Text Understanding Competence. A Case Study: POLINT-112-SMS, *Proc. of the LRPS Workshop at LREC 2012, May 27, 2012. Istanbul, Turkey*, ELRA: Paris, 54-63.

Vetulani, Z. (2014): PolNet - Polish WordNet, in: Z. Vetulani, and J. Mariani (Eds.): *Human Language Technology. Challenges for Computer Science and Linguistics. LTC 2011. Revised Selected Papers.* LNAI 8387, Springer-Verlag Berlin Heidelberg, 408-416.

Vetulani, Z., Vetulani, G. (2014a): Through Wordnet to Lexicon Grammar, in. F. Kakoyianni Doa (Ed.). *Penser le lexique grammaire: perspectives actuelles*, Editions Honoré Champion, Paris, 531-543

Vetulani, Z., Vetulani G. (2014b): Verb-Noun Collocations in PolNet 2.0. In V. Henrich and E. Hinrichs (Eds.): *Proceedings of the Workshop on Computational Cognitive, and Linguistic Approaches to the Analysis of Complex Words and Collocations (CCLCC 2014)* Tübingen, Germany, 73-77.

Vetulani, Z., Vetulani, G. (2015): Synonymie et granularité dans les bases lexicales du type Wordnet, *Studia Romanica Posnaniensia,* Vol. 42/1, Wydawnicywo Naukowe UAM, Poznań, 113-127.

Vetulani, Z., Vetulani, G., Kochanowski, B. (2016): Recent Advances in Development of a Lexicon-Grammar of Polish: PolNet 3.0. In N. Calzolari et al. (Eds.), *The 10th Conference on Language Resources and Evaluation*, ELRA, Paris, France, 2851-2854.

Vetulani, Z., Osiński, J. (2017): Intelligent Information Bypass for More Efficient Emergency Management", *Computational Methods in Science and Technology* 23(2), 105–123.

Vossen,P., Bloksma, L., Rodriguez, H., Climent, S., Calzolari, N., Peters, W. (1998): *The EuroWordNet Base Concepts and Top Ontology, Version 2, Final, January 22, 1998 (Euro WordNet project report).*

Proceedings, pages 393–397
Paris, UNESCO Headquarters, 5-6 December, 2019.

Analysis of Language Relatedness for the Development of Multilingual Automatic Speech Recognition for Ethiopian Languages

Martha Yifiru Tachbelie[1,2], Solomon Teferra Abate[1,2], Tanja Schultz[1]
[1]Cognitive Systems Lab
University of Bremen, Germany
[2]School of Information Science
Addis Ababa University, Ethiopia
{marthayifiru, abate, tanja.schultz}@uni-bremen.de

Abstract

In this paper, we present the analysis of GlobalPhone (GP) and speech corpora of Ethiopian languages (Amharic, Tigrigna, Oromo and Wolaytta). The aim of the analysis is to select speech data from GP for the development of multilingual Automatic Speech Recognition (ASR) system for the Ethiopian languages. To this end, phonetic overlaps among GP and Ethiopian languages have been analyzed. The result of our analysis shows that there is much phonetic overlap among Ethiopian languages although they are from three different language families. From GP, Turkish, Uyghur and Croatian are found to have much overlap with the Ethiopian languages. Moreover, morphological complexity of the GP and Ethiopian languages, reflected by type to token ration (TTR) and out of vocabulary (OOV) rate, has been analyzed. Korean and Amharic have been identified as extremely morphologically complex compared to the other languages. Tigrigna, Russian, Turkish and Polish are also among the morphologically complex languages.

Keywords: Language relatedness, Multilingual ASR, GlobalPhone, Ethiopian Languages

አጠቃሎ

በዚህ ፅሑፍ የምናቀርብላችሁ በግሎባል ፎን እና በአራት የኢትዮጵያ ቋንቋዎች (አማርኛ፣ ትግርኛ፣ ኦሮምኛ እና ወላይትኛ) የድምፅ ዳታ መካከል ያደረግነውን የማነጻጸር ጥናት ነው። የጥናቱ ዋና አላማ ከግሎባል ፎን የድምፅ ዳታ ውስጥ ለኢትዮጵያ ቋንቋዎች ንግግርን ወደ ድምፅ የሚቀይር መተግበሪያ ለመስራት ጠቃሚ የሆነ የድምፅ ዳታ መምረጥ ነው። ለዚህም በቋንቋዎቹ መካከል ያለውን የድምፅ መመሳሰል አጥንተናል። ውጤቱ እንደሚያሳየው ምንም እንኳን በተለያዩ የቋንቋ ዝርያ ውስጥ ቢሆኑም፣ አራቱም የኢትዮጵያ ቋንቋዎች በጣም ከፍተኛ የድምፅ መመሳሰል አላቸው። ከግሎባል ፎን ቋንቋዎች መካከልም ቱሪክኛ፣ ክሮኤሽያኛ እና ኡይገርኛ ከኢትዮጵያ ቋንቋዎች ጋር ብዙ የድምፅ መመሳሰል አሳይተዋል። በተጨማሪም የቋንቋዎቹን ምእላዳዊ ውስብስብነት ለመረዳት እንዲቻል "Out of Vocabulary" እና "Type to Token Ratio" አስልተናል። አማርኛ እና ኮሪያኛ በጣም ከፍተኛ ምእላዳዊ ውስብስብነት እንዳላቸው ውጤቱ አሳይቷል። ትግርኛ፣ ሩሲያኛ፣ ቱርክኛ እና ፖለንኛም ከፍተኛ ምእላዳዊ ውስብስብነት ካላቸው ቋንቋዎች መካከል ናቸው።

1. Introduction

With more than 7000 languages in the world (Ethnologue, 2019) and the need to support multiple input and output languages, it is one of the most pressing challenge for the speech and language community to develop and deploy speech processing systems in yet unsupported languages rapidly and at reasonable costs (Schultz, 2004; Schultz and Kirchhoff, 2006). Major bottlenecks are the sparseness of speech and text data with corresponding pronunciation dictionaries, the lack of language conventions, and the gap between technology and language expertise. Data sparseness is a critical issue due to the fact that speech technologies heavily rely on statistical modeling schemes, such as Hidden Markov Models, Deep Neural Networks (DNN) for acoustic modeling and n-gram and DNN for language modeling. Although statistical modeling algorithms are mostly language independent and proved to work well for a variety of languages, reliable parameter estimation requires vast amounts of training data. On the other hand, large-scale data resources for research are available for less than 100 languages and the costs for these collections are prohibitive to all but the most widely spoken and economically viable languages. This calls for the development of cross-lingual and/or multilingual speech processing/recognition systems.

In cross-lingual speech recognition, resources of a language (source/donor language) are used to develop speech recognition system for another (target language) with or without little adaptation data from the target language. Multilingual speech recognition system is described as a system that is able to recognize multiple languages which are presented during training(Schultz and Waibel, 2001). (Vu et al., 2014) described multilingual ASR as system in which at least one of the components (feature extraction, acoustic model, pronunciation dictionary, or language model) is developed using data from many different languages. Although multilingual ASR systems are useful in other contexts, they are particularly interesting for under-resourced languages where training data are sparse or not available at all(Schultz and Waibel, 2001). Furthermore, they provide an appealing solution for multilingual, multi-Ethnic, and economically disadvantaged countries, such as Ethiopia.

Ethiopia is a multilingual and multi-ethnic country where over 80 languages are spoken by the citizens. (Ethnologue, 2019) states that, “The number of individual languages listed for Ethiopia is 90. Of these, 88 are living and 2 are extinct. Of the living languages, 85 are indigenous and 3 are non-indigenous.“

When it comes to language resources required for the development of speech and language processing tools, almost all Ethiopian languages are under-resourced. On the other hand, developing large-scale language resources is not economically viable. Thus, alternative approaches need to be used to make Ethiopians benefit from speech and language processing tools. Accordingly, we are currently investigating the development of multilingual speech recognition system for Ethiopian languages. For this purpose, we will use GlobalPhone(Schultz et al., 2013), a multilingual database of high-quality read speech with corresponding transcriptions and pronunciation dictionaries in more than 20 languages. To use this resource, it is mandatory to identify which languages are closely related with Ethiopian languages and therefore will be useful in the development of multilingual ASR.

In this paper, we present the analysis of GlobalPhone (GP) and speech corpora of Ethiopian languages (Amharic, Tigrigna, Oromo and Wolaytta) which are recently developed(Abate et al., 2020). The aim is to select speech data from GP and related Ethiopian Languages for the development of multilingual Automatic Speech Recognition (ASR) system for the Ethiopian languages. We have analysed the phonetic overlaps among GP and Ethiopian languages based on International Phonetic Association (IPA) sound representation. Moreover, morphological complexity of the GP and Ethiopian languages has been analyzed based on type to token ration (TTR) and Out of Vocabulary (OOV) rate calculated on the basis of training transcriptions.

2. Available Resources

2.1. GlobalPhone

GlobalPhone (GP) is a multilingual data corpus that comprises: 1) audio/speech data, i.e. high-quality recordings of spoken utterances read by native speakers, 2) corresponding transcriptions, 3) pronunciation dictionaries covering the vocabulary of the transcripts, and 4) baseline n-gram language models. The first two are referred to as GP Speech and Text Database (GP-ST), the third as GP Dictionaries (GP-Dict), and the latter as GP Language Models (GP-LM). GP-ST is distributed under a research or commercial license by two authorized distributors, the European Language Resources Association (ELRA) (ELRA, 2012) and Appen Butler Hill Pty Ltd. (Ltd, 2012). GP-Dict is distributed by ELRA, while the GP-LMs are freely available for download from our website (LM-BM, 2015).

The entire GP corpus provides a multilingual database of word-level transcribed high-quality speech for the development and evaluation of large vocabulary speech processing systems in the most widespread languages of the world. GP is designed to be uniform across languages with respect to the amount of data per language, the audio quality (microphone, noise, channel), the collection scenario (task, setup, speaking style), as well as the transcription and phone set conventions (IPA-based naming of phones in all pronunciation dictionaries). Thus, GP supplies an excellent basis for research in the areas of (1) multilingual ASR, (2) rapid deployment of speech processing systems to yet unsupported languages, (3) language identification tasks, (4) speaker recognition in multiple languages, (5) multilingual speech synthesis, as well as (6) monolingual ASR. The GP corpus covers 20 languages, i.e. Arabic (modern standard), Bulgarian, Chinese (Mandarin and Shanghai), Croatian, Czech, French, German, Hausa, Japanese, Korean, Polish, Portuguese, Russian, Spanish, Swedish, Tamil, Thai, Turkish, Ukrainian, and Vietnamese. It comprises wide-spread languages (e.g. Arabic, Chinese, Spanish, Russian), contains economically and politically important languages, and spans wide geographical areas.

In addition to GP languages, we have also considered Uyghur for which we have a speech corpus and want to use it in multilingual setting. The Uyghur corpus is a read speech corpus of selected newspaper articles. It contains 12 hours of training speech collected from 41 speakers with 4k sentences and 1.5 hours of evaluation speech collected from 5 speakers with 491 utterances.

2.2. Speech Corpora of Ethiopian Languages

Speech corpora of four Ethiopian Languages (Amharic, Tigrigna, Oromo and Wolaytta) are considered in the analysis. Two Amharic speech corpora are considered. 1) The Amharic read speech corpus prepared at the University of Hamburg (Abate et al., 2005), referred as AM2005. This contains 20 hours of training speech collected from 100 speakers who read a total of 11k sentences, development and test sets read by 20 other speakers (10 each). 2) The Amharic read speech corpus, referred as AM2020, prepared at Addis Ababa University (AAU) together with the preparation of speech corpora of the other three languages.

The corpora of the Ethiopian languages have been collected in Ethiopia under a thematic research funded by AAU (Abate et al., 2020). The Amharic, Tigrigna and Oromo speech corpora consist of speech of 98 readers each. Most of the speakers of the languages, read 121 to 130 sentences. The size of the training speech of these three languages is 26 hours for Amharic and 22 hours for Tigrigna and Oromo. The Wolaytta corpus consists of recordings of 85 speakers where most of them read 140-150 sentences. Considering the difficulty of getting 100 readers for Wolaytta and aiming at collecting not less than 20hrs of speech, 150 utterances were assigned to each speaker. This way it became possible to collect a speech corpus of 29 hours.

3. Analysis Of Globalphone and Ethiopian Languages Corpora

3.1. Language Family

The languages considered in our analysis fall into 10 language families. Austro-Asiatic: Hausa and Vietnamese; Cushitic: Oromo; Indo-European: that includes Germanic (English, German and Swedish), Ro-

mance (French, Portuguese and Spanish) and Slavic (Bulgarian, Croatian, Czech, Polish and Russian); Japonic: Japanese; Koreanic: Korean; Kra-Dai: Thai; Omotic: Wolaytta; Semitic: Amharic, Arabic and Tigrigna; Sino-Tibetan: Mandarin; Turkic: Turkish and Uyghur.

3.2. Writing System

The written language contains all types of writing systems, i.e. logographic scripts (Chinese Hanzi and Japanese Kanji), phonographic segmental scripts (Roman, Cyrillic), phonographic consonantal scripts (Arabic), phonographic syllabic scripts (Japanese Kana, Thai), abugida/Ethiopic (Amharic, Tigrigna), linear nonfeatural (Uyghur) and phonographic featural scripts (Korean Hangul).

3.3. Sound System

Phonetic information is important in multilingual speech recognition. Considering this fact, we have analysed the sound system of the GP as well as Ethiopian languages. In the anaysis, a broad selection of phonetic characteristics have been considered, e.g. tonal sounds (Mandarin, Thai, Vietnamese, Oromo, Wolaytta), consonantal clusters (German), nasals (French, Portuguese), plosive sounds (Amharic, Oromo, Tigrigna, Wolaytta), uvular (Uyghur) and palatized sounds (Amharic, Oromo, Tigrigna, Wolaytta, Russian).

No. of Lang.	No. of Phone	Consonants	Vowels
All	3	m, n, s	-
22	3	k, l, t	-
21	5	b, d, f, p	u
20	1	j	-
19	3	g	i, o
17	2	v, z	-
16	3	h,ʃ	a
15	2	r	e
13	2	w	ɛ
12	4	x,ɲ, ʧ, ʒ	-
11	1	ŋ	-
9	5	ʤ	aː, iː, ə ɨ
8	1	-	ɔ
7	5	ʦ, ʔ	oː, uː, ʌ
6	2	ʨ	eː
5	6	ʎ, ʁ, t^h	y, ai, ø
4	13	c,ç,k^h,kʼ, p^h, pʼ, ɾ, ʂ tʼ, ʧʼ, ʐ	ɪ, uə
3	18	ɖ, ð, ɣ, sʼ, s^j, t^j d͡z	ɛː, œ, ɯ, au, aʊ, eɪ, ja, ju, oʊ, ɔɪ, ua
2	39	b^j, d^j, ɟ, ħ, ʕ, ɦ, m^j, p^j, ɑ, ɕ, ʈ, ʈ͡ʂ, ʨ^h, θ, d͡z, v^j, z^j, ʑ	aˌ, ɐ, æ, é, əː, ɛ̃, iˌ, í, oˌ, ɔ̃, øː, uˌ, ʊ, yː ɯː, iə, je, jo, uɨ, ui, ue

Table 1: Polyphones shared by 2 or more languages

We have analyzed and identified language independent phones (polyphones), phones occurring in more than one languages, and language dependent (monophones), phones that occur in only one language(Andersen et al., 1993) as it is done in(Schultz and Kirchhoff, 2006). The phone analysis is done on the basis of the pronunciation dictionaries we have at hand for each of the languages. Table 1 indicates the polyphones that occur in two or more languages.

In addition to identification of polyphone and monophone, we analyzed the sound overlap among the GP and the four Ethiopian languages based on the sound representation of the International Phonetic Association (IPA). That means we considered sounds from different languages similar if they are represented by the same IPA symbol. Otherwise, they are considered as different sounds. Figure 1 indicates the coverage of sounds of a language (values being 0% to 100%), for example Amharic, in the rest of the languages. The dark blue color indicates 100% overlap whereas light yellow indicates low or no overlap.

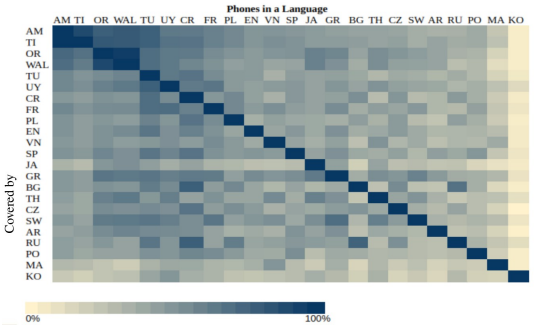


Figure 1: The phonetic sound overlap.

As can be seen from Figure 1, there is much phone overlap among the Ethiopian languages. Interestingly, all Amharic phones are covered by Tigrigna. We expected high phonetic overlap among the Semitic languages (Amharic, Tigrigna and Arabic), however Arabic has low phone overlap with the rest of the Semitic languages. From GP, Turkish, Uyghur and Croatian cover most phones of the Ethiopian languages. The Figure also shows that there is high phone overlap among three of the Slavic languages: Croatian, Bulgarian and Russian. Although, Polish and Czech also fall under Slavic language family, our analysis does not show high phone overlap of these languages with the rest of the Slavic languages. On the other hand, Korean and Mandarin seem to have less phone overlap with the rest of the GP as well as Ethiopian languages.

3.4. Morphological Property

The morphological complexity of a language affects the quality of a language model and the coverage of decoding vocabulary (pronunciation dictionary). Since language model and pronunciation dictionary are components of a speech recognition system that affect performance, we have analyzed the morphological complexity of the GP and Ethiopian languages based on the training and the evaluation set transcriptions.

The languages considered in our analysis cover many morphological variations, e.g. agglutinative lan-

guages (Turkish, Korean), compounding languages (German), and non-concatenative root-pattern morphology (Amharic, Tigrigna and Arabic), and also include scripts that completely lack word segmentation (Chinese, Thai).

We have computed type-token ratio (TTR), calculated as vocabulary size divided by text length, based on the training transcriptions. (Kettunen, 2014) showed that TTR can order the languages quite meaningfully in a morphological complexity order or at least groups most of the languages with same kind of morphological complexity and clearly separates the most and least morphologically complex languages. Since TTR is affected by the length of the text sample, we also computed average TTR (ATTR) for the training transcription based on k disjunct 1000-utterance subsets of the training set. In each iteration we randomly select 1000 utterances, compute TTR, remove these utterances from the pool, and then continue with the next iteration. Finally, the ATTR is computed using TTR values of each distinct 1000 utterances. Figure 2 and 3 show the TTR and ATTR for each of the languages considered in our analysis.

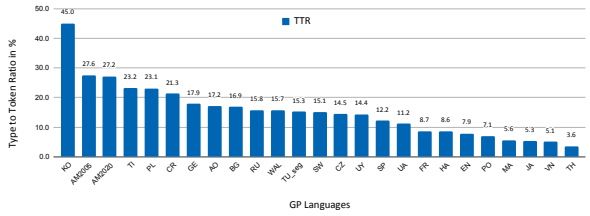


Figure 2: Type-Token Ratio.

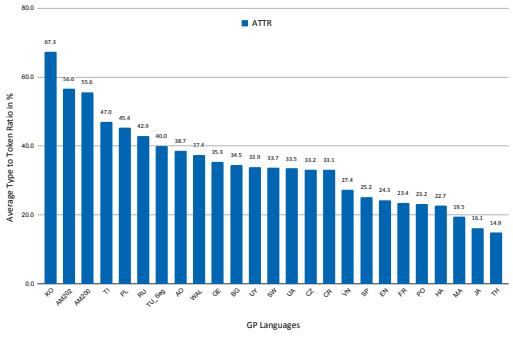


Figure 3: Average Type-Token Ratio.

As shown in both Figures, TTR and ATTR reflect the morphological complexity of the languages. The languages are arranged from morphologically more complex to less complex. Korean is indicated as the most morphological complex language, followed by Amharic (for both Amharic corpora) and Tigrigna. On the other hand, English, Hausa, Japanese, Mandarin and Thai are identified as less morphological complex languages. However, we used segmented at word (multi-character or multi-syllable) transcription for Japanese, Mandarin and Thai and therefore, the Figure may not reflect the true morphological property of these languages. But, in general, the ATTR seem to reflect the morphological complexity of the languages.

Another metric that reflects morphological complexity and commonly used in the ASR community is Out of Vocabulary (OOV) rate. In ASR, one OOV word accounts to one or one and half wrongly recognized word/s. Mostly high OOV means high word recognition error rate. We have calculated the OOV rate (see Figure 4) of the evaluation set of each of the languages against the different vocabulary sizes taken from the training transcriptions. Our analysis of the morphological complexity of the GP and Ethiopian Languages, both using TTR and OOV, helps to know which languages are more challenging with respect to the two components of the ASR system: vocabulary and language models. As a solution to the morphological complexity problem, morphemes (instead of words) have been used as units in these models. The other alternative, i.e. the use of large vocabularies and language models are limited by the (very) small amounts of data available for under-resource languages. Depending on the availability of resources, we will study the impact of these approaches in multilingual as well as monolingual ASR of morphologically complex languages.

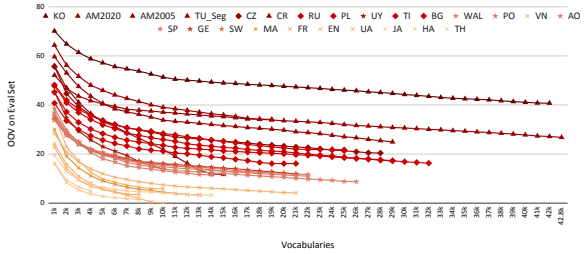


Figure 4: OOV of the Evaluation Set.

4. Conclusion and Future Direction

In this paper we presented the analysis of GlobalPhone and four Ethiopian languages speech corpora that we intend to use for the development of multilingual speech recognition system for Ethiopian languages. The purpose of the analysis is to select viable speech corpora from the available resources. The phonetic analysis shows that there is high phone overlap among Ethiopian languages although the languages are from three different language families. Moreover, Turkish, Uyghur and Croatian are found to have slightly high phone overlap with the Ethiopian languages. Our analysis also shows that, the Ethiopian languages', considered in the analysis, morphology is not simple. Amharic has extremely high morphological complexity, next to Korean.

Our next step is conducting multilingual ASR experiments by adding Turkish, Uyghur and Croatian speech corpora from GP to the Ethiopian languages corpora.

5. Acknowledgments

The first two authors are grateful to Alexander von Humboldt for the experienced researchers fellowship grant.

6. Bibliographical References

Abate, S. T., Menzel, W., and Tafila, B. (2005). An amharic speech corpus for large vocabulary continuous speech recognition. In *INTERSPEECH*.

Abate, S. T., Melese, M., Tachbelie, M. Y., Meshesha, M., Atinafu, S., Mulugeta, W., Assabie, Y., Abera, H., Ephrem, B., Abebe, T., Tsegaye, W., Lemma, A., Andargie, T., and Shifaw, S. (2020). Speech corpora for four ethiopian languages: Amharic, oromo, tigrigna and wolaytta. In *LREC2020*.

Andersen, O., Dalsgaard, P., and Barry, W. J. (1993). Data-driven identification of poly- and mono-phonemes for four european languages. In *EUROSPEECH*.

ELRA. (2012). European language resources association elra. ELRA catalogue. Retrieved November 30, 2012, from http://catalog.elra.info.

Ethnologue. (2019). Languages of the world. Retrieved October 21, 2019, from https://www.ethnologue.com/.

Kettunen, K. (2014). Can type-token ratio be used to show morphological complexity of languages? *Journal of Quantitative Linguistics*, 21(3):223–245.

LM-BM. (2015). Benchmark globalphone language models. Retrieved October 21, 2019, from https://www.csl.uni-bremen.de/GlobalPhone/.

Ltd, A. B. H. P. (2012). Speech and language resources 2012. Appen Butler Hill Speech and Language Resources 2012 - Product Catalogue.

Schultz, T. and Kirchhoff, K. (2006). *Multilingual Speech Processing*. Elsevier Academic Press.

Schultz, T. and Waibel, A. (2001). Language-independent and language-adaptive acoustic modeling for speech recognition. *Speech Commun.*, 35(1-2):31–51, August.

Schultz, T., Vu, N. T., and Schlippe, T. (2013). Globalphone: A multilingual text and speech database in 20 languages. In *ICASSP*.

Schultz, T. (2004). Towards rapid language portability of speech processing systems. In *Conference on Speech and Language Systems for Human Communication (SPLASH)*, volume 1, Delhi, India, November.

Vu, N. T., Imseng, D., Povey, D., Motlícek, P., Schultz, T., and Bourlard, H. (2014). Multilingual deep neural network based acoustic modeling for rapid language adaptation. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7639–7643.

Proceedings, pages 398–401
Paris, UNESCO Headquarters, 5-6 December, 2019.

Increasing Diversity via Augmented and Distributed Online Conferences

Alejandrina Cristia
Laboratoire de Sciences Cognitives et de Psycholinguistique,
Département d'études cognitives, ENS, EHESS, CNRS, PSL University
29 rue d'Ulm, 75005, Paris, France
alecristia@gmail.com

Abstract

Language and speech technologies can become available 'for all' only if we increase access to knowledge and networking opportunities, particularly for scientists who are geographically distant from Europe and North America, and who likely have a limited travel budget. I propose a relatively cheap and scalable strategy to do so: The creation of augmented and distributed online conferences. Participants travel to a central areal location to view the streamed conference together, and online (video-)chats allows such virtual attendees to interact with physical conference attendees. Benefits include the promotion of areal networks, increased learning through active participation, and reduced environmental costs.

Keywords: knowledge dissemination, diversity, networking, environmental impact

Résumé

Las tecnologías de lenguaje y habla pueden estar disponibles 'para todos' solo si aumentamos el acceso al conocimiento y las oportunidades de establecer contactos, particularmente para los científicos que están geográficamente distantes de Europa y América del Norte, y que probablemente tienen un presupuesto de viaje limitado. Propongo una estrategia relativamente barata y escalable para hacerlo: la creación de conferencias en línea distribuidas y aumentadas. Los participantes viajan a una ubicación central para ver la conferencia transmitida juntos, y (video-)chats en línea permiten que dichos asistentes virtuales interactúen con los asistentes físicos a la conferencia. Los beneficios incluyen la promoción de redes de área, mayor aprendizaje a través de la participación activa y costos ambientales reducidos.

1. Motivation

In speech and language technology, a primary mode of knowledge creation and dissemination are conferences. In this paper, I will set aside the issue of whether and why paper evaluation should be tied together with a physical presentation. Instead, I will focus on short-comings of current conference organization in terms of geographic and linguistic biases, and some related problems.

First, access to conferences is easier for scientists who are located close to where conferences because of lower time and moneyf travel costs. By and large, this advantages scientists in the Northern Hemisphere, and typically those residing in North America and Europe, thus disadvantaging researchers from low- and- middle-income (LMIC) countries and even scientists living in relatively wealthy countries but with no travel budget. Sometimes, researchers attempt to counter this bias by convening conferences in other locations. When doing so, however, they worsen the second issue, as follows.

Second, physical conferences imply a steep cost to the environment. In a recent study of six conferences from the European Consortium for Political Research (which thus was attended primarily by attendees from the same continent), (Jäckle, 2019) estimated each visitor caused .5-1.3 tons of CO_2 equivalents per conference when they happened in Europe, and 1.9-3.4 tons for one held in Montreal.

Third, even in cases where scientists from LMIC can access conferences, they likely benefit less. To begin with, very often conferences are conducted in a major language which may not be those scientists' native tongue, meaning they may learn less when viewing presentations because they cannot adjust the rhythm or speed of the presentation. Moreover, question periods may happen too quickly for someone who may need a moment to find their words when speaking. As a result, such attendees may get used to the idea of not asking questions, becoming passive listeners. In addition to this having a negative impact in their informational uptake (Jensen et al., 2015), this deprives them from an opportunity to display themselves and attract colleagues' attention.

This connects with a fourth issue that is likely to result in decreased diversity in academia, which is inequality in access to networking opportunities, for which conferences appear irreplaceable. Giving a presentation or even asking a question that others shared is an excellent conversation starter, as these other audience members may come up to the presenter or question asker to continue the conversation. Coffee breaks and other open discussion periods are often prized moments to approach researchers with whom we have a passing acquaintance that we want to reinforce, or even people we would like to introduce ourselves to. However, even if scientists from LMIC attend conferences, these times may be yet another stressful opportunity, when one is forced to socialize orally in a non-native language now in the presence of background noise high enough to make even native speakers find the experience tiresome. Over time, attempting to follow presentations and discussions in a non-native language causes fatigue, which compounds all of the aforementioned negative effects.

1.1. Current alternatives

In the scientific world, there have been repeated calls to move towards online conferences, primarily in order to minimize their environmental impact. For example, a re-

cent proposal has been made for "nearly carbon-neutral conferences", based on several successful attempts (Ken Hiltner, 2017), and which assumes that physical presence may no longer be required even for presenters because they can send in video-recordings of their presentations, and discussion can occur in the form of comments to blog entries. This is certainly a promising direction, and could eventually fully replace physical conferences. However, many will find such an idea threatening, requiring a whole rethinking of the place of conferences in today's professional world. Additionally, it is unclear whether the same level of engagement and comprehension can be obtained when there is no real-time interaction.

Twitter users may feel that there is already a possibility for people not attending a conference to follow it, provided someone is live tweeting the event. Others have already pointed out some limitations of this, including most saliently that the short format required by twitter virtually enforces simplification (Erin Zimmerman, July 30 2019). In addition, this system is mainly for individuals who are physically present to broadcast what they choose to discuss, with limited opportunities for researchers in LMIC to impact the discussion. Additionally, there is no structure supporting networking for these peripheral participants.

Several of the problems summarized above relate to processing and producing non-native speech. In the industrial world, there are current solutions for the question of language mismatches between attendee and speaker. For instance, Interactio provides online translation services that are streamed directly to audience members' smartphones (**?**)interactio). I think it is unlikely such a solution will work for science because of its economic cost and the difficulty of finding translators who can follow and accurately translate scientific presentations.

2. Augmented and Distributed Online Conferences (ADOC)

The main idea is to augment the current system of physical, centralized conferences using online resources and geographic distribution. I will discuss three use cases, the first being when there is a conference that has online streaming; the second a conference with no online streaming; and the third for a conference being organized from scratch.

2.1. Augmenting a conference that has online streaming

Let us assume there exists a conference whose conveners have decided to live stream. Even without discussion with and agreement by the conveners, the audience can self-organize to increase the chances that potential audience members who cannot attend physically nonetheless benefit from the content presented, live discussions, and some networking opportunities.

The first component of an ADOC is its augmentation, and I propose it can currently rest on private but free platforms: Slack, Google products (Forms, Hangouts), and Youtube. A Slack workspace should be created with the following features:

- An Announcements channel, where key information is posted (including descriptions of all channels, and instructions for creating new channels)
- A channel for each session, so that (e-)attendees to that session can post questions/comments on it
- A bio channel, where groups and/or individuals can post some information about themselves, including keywords representing their research interest

This workspace should be widely advertised and a google form should be created. This form serves to collect email addresses necessary to invite respondents to the Slack workspace. In addition, the form asks them to specify whether they'll be attending the event physically or online. In the former case, respondents are asked whether they are open to being the question askers for online participants for a given session (and if so, which). Respondents are also asked whether they volunteer to help with the organization of the ADOC. Respondents who participate online and agree to volunteer can help manage the Slack workspace (adding participants, trouble-shooting, moderating channels when needed).

The second component of a ADOC is its distributed nature, whereby events that coinicide in time with the actual conference are organized in geographically distributed locations. Let us imagine that three people in Senegal would like to e-attend an event in Paris. They would book a room in one of their universities, which has a good bandwidth and online conferencing equipment (a camera, a microphone, a large screen). They would also look into opportunities to have food delivered, or instead decide to skip connecting with the Parisian event during lunch breaks. They would either agree to all watch the same event on the large screen, with an opportunity to discuss it; or instead decide to watch different events, and recap the results for each other. Let us imagine that a group of seven people in Mali will be doing the same.

As these online attendees follow the conference, they may ask questions on the Slack channel dedicated to each session. Other physical or online attendees may choose to answer the question, or +1 the question to indicate they would like for it to be asked in the question period. When question period comes, physical attendees who agreed to be question askers would ask the most up-voted question.

Throughout the conference, online and physical attendees can also create new channels to continue targeted discussions. Imagine someone asks a question about the precise structure of an RNN to be used in a diarization task, which he/she knows interests two other people. They can create a channel called "RNN-dia" a google hangouts room and advertise this in the Announcements channel. After the conference is over, a longer term collaboration could emerge between these eight people if they continue the discussion within the Slack workspace, through contact information shared that day, or present in their bios (in the bio channel). In addition, Slack has a system for direct messaging. Although not the same as standing in line to introduce oneself to someone famous, this would allow e-attendees a chance to connect with these people (provided both sign up to the e-conference).

2.2. Augmenting a conference that does not have online streaming

This case is almost exactly as the above, except that the physical volunteers become crucial because not only will they ask questions in the question period, but they should also video- or at least audio-stream the conference for the people who are e-attending. To this end, they must create an account on Youtube, and learn how to live stream to a private channel. It must be to a private channel because attendees to the conference did not sign a release for their image and speech to be shared publicly with no restraint. In addition to this change, the Google Form in which people sign up to attend the e-conference must include an ethics section explaining that it is not allowed to share the Youtube channel link with third parties, or to reproduce the content elsewhere.

2.3. Creating a conference from scratch

For readers who are creating a conference from scratch, they could borrow ideas from (Ken Hiltner, 2017) and make sure they provide people in the periphery an equal chance for being speakers. In addition, they can make sure the event is live-streamed, using the recommendations in that white paper (which build on Youtube). Ideally, the costs for running Slack on a private server and making all other services private would be included in the conference registration, so that the data exchanged does not automatically go to Slack/Google.

Also, it would be ideal for such conference organizers to make sure that there is appropriate space, time, and bandwidth so that physical attendees can interact with the e-attendees during coffee breaks and poster sessions. For instance, there could be alcoves or other smaller spaces ideal to have Google hangout discussions using one's phone. Even better, one could fit these spaces with tablets and headsets, directly linked to the Slack/Google environment.

3. Comparison between DACO and conferences as usual

3.1. Expected relative benefits

Numerous relative benefits are expected. To begin with, being with others (in the distributed setting) or at least connected online with physical attendees means that online attendees would benefit from a more active experience, so they may learn better during the talks. In addition, non-native listeners who meet up with others in the distributed setting may be able to discuss and translate into their native language, further improving understanding.

Questions in the question period would represent people who are not physically in the audience. This benefits not only e-attendees, whose voice may now be heard, but also presenters, who can benefit from a wider diversity of perspectives on their work. Indeed, I suspect that the discussions in the session-specific channels will be beneficial to both e- and physical attendees.

E-attendees would benefit from increased chances to network with others present in the conference, a possibility that is completely absent in the current system. In fact, it is often difficult to connect with others having similar interests in large conferences, simply because there are too many people and there is typically nothing in their physical appearance that allows us to know we share an interest. Moving some of these issues to the virtual world may further facilitate discussion with others even for physical participants.

In today's conferences, fascinating discussions that one strikes in the coffee break are quickly forgotten as one speeds to the next talk. By creating a structured space to carry on targeted conversations, ADOCs allow a physical trace of some of these conversations, and may even provide a natural setting for these ephemeral chats to flourish into collaborations.

There are additionally two benefits that are not related to allowing participation from geographically and linguistically diverse audiences. First and foremost, ADOCs would allow some to optionally not attend physically, resulting in reduced environmental impact of the conference. Additionally, they may increase diversity in other ways, being a good option for primary caregivers who cannot physically travel far or for an extended period of time; they may provide unique interaction opportunities to people who are shy or suffer from social anxiety; and they also benefit students and junior researchers, who typically do not have a travel budget and cannot afford physical participation to conferences.

3.2. Expected relative costs

The main cost is time: to organize the workspaces, write instructions, publicize, process incoming information, train volunteers. This should be measured, and could be compared against the time savings of reduced travel.

In addition, there may be some security risks. Particularly in conferences that do not have a live-streaming component, some may feel challenged about having their talks live-streamed to a private Youtube channel. In general, today's attendees may operate under the assumption that presentations and comments are done in a closed room and will be quickly forgotten. Thus, a shift in how speakers present may be required.

3.3. Measurement of effects

While writing this paper, I realized that although there are many discussions regarding the role of conferences, their importance for career development, their negative environmental effects, there seem to be relatively few actual quantitative studies teasing apart correlation from causation, and opinion from fact (e.g., (Prpić, 2000; Rowe, 2018)). Work on actually proposing alternatives and measuring relative impact is, to my knowledge, non-existent.

It is crucial to actually calculate the benefits and costs of different types of conferences or conference features from now on. Often the opening talk at a conference comments on diversity in terms of the proportion of people coming from different sites, which seems a good place to start, but an insufficient metric.

We could measure comprehension of talks by asking speakers to use online polls to check for content comprehension, providing different links to online versus physical attendees; the prediction is that e-attendees who participate in distributed groups understood the talk's contents better than

those watching alone; and a key question is whether those two groups understood the talk's contents any differently from physical attendees.

Using names or other identifiers provided during registration could help direct network analyses. One prediction is that physical conference attendees and e-attendees are more likely to cite each other, or co-publish, after the conference than before; a control group can also be defined using a conference without ADOC, to calculate the base rate of mutual citation. It is an open question whether physical attendees will cite each other more than they cite e-attendees, and vice versa. We can also test the prediction is that post-doctoral researchers and students who e-participate are more likely to be hired by e- or physical participants than a group of control students, or some established base rate, with follow-up hypotheses regarding the relative advantage of students/post-docs who attended physically versus online only.

Finally, one could check the prediction is that e-attendees will have much shorter travel times, even if they meet a geographically distributed group, and will contribute less C02 and other environmental costs. Another key question is whether e- and physical attendees end up attending a similar proportion of the talks.

3.4. Limitations

While ADOC may provide a solution to some of the problems current purely physical, centralized conferences pose, they do not fix all of them. In particular, this solution is not useful to scientists living in countries where the internet infrastructure is poor. Additionally, they represent only a partial solution for those who cannot book a well-equipped videoconferencing room in their university, and instead may need to log in from a home or a café. Additionally, some conference organizers may feel they are losing money in this way, and it is unfair for e-attendees to not pay any of the costs the conference conveners have engaged in.

Perhaps in the future conferences will be organized with these three limitations in mind. They could be addressed by organizing the conference as a distributed event from scratch, such that the fee depends on which site the attendee goes to. If attending from UK, the fee may be around 500$ taking into account not only the costs of the central organization of the program but also the conference center, food delivery, and other costs local to the UK; but attendees to the Buenos Aires site may only pay 100$ for the same services. This fee would thus pay for the local videoconference and food costs in an actual conference site, whereas some of it would go to the central conference conveners.

Even in this case, however, some issues remain unresolved. We still do not have a good solution for the fact that publication in major languages still benefits native speakers of those major languages, whereas publication in smaller languages currently has a much lower readership. Similarly, none of these solutions are perfect to balance the scale in terms of networking: Scientists from famous labs may still find it easier to meet people, find good collaborators, hire and be hired, than scientists in less prominent sites.

3.5. LT4All as a use case

The International Conference Language Technology for All provides an excellent use case for the ADOC proposal, falling neatly in the first use case mentioned above (a conference that is live streamed). I have set up a Slack workspace (`lt4all2019.slack.com`) and a Google Form (`https://forms.gle/yTT88GR1nWkWdNs8A`).

This paper is available as a preprint from `https://osf.io/s4bta`. In that version, I may be able to update the information on some of the metrics proposed above. Geographical origin of physical and e-attendees, as well as travel time, will be gathered via an anonymized Google Form, and can serve to estimate C02 and other environmental costs. If I find any speakers who are willing to use online polls to check for content comprehension, this could serve to answer whether distributed and/or online participation can be equivalent in terms of learned content.

4. Conclusion

In this paper, I have proposed the concept of Augmented and Distributed Online Conferences (ADOCs), as a way to increase the dissemination of scientific information that is available in conferences to scientists who may reside in geographically spread locations, with likely added benefits in terms of networking opportunities. Ultimately, the question of how useful ADOCs are is an empirical one, and I look forward to there being data to answer it.

5. Acknowledgements

I am grateful to the womenNspeech mailing list and Emmanuel Dupoux for helpful discussion. All errors are my own.

6. References

Erin Zimmerman. (July 30, 2019). Live tweeting science conferences: The good, the bad, and the ugly. last visited 2019-11-20.

Jäckle, S. (2019). WE have to change! The carbon footprint of ECPR general conferences and ways to reduce it. *European Political Science*, 18:630–650.

Jensen, J. L., Kummer, T. A., and Godoy, P. D. d. M. (2015). Improvements from a flipped classroom may simply be the fruits of active learning. *CBE–Life Sciences Education*, 14(1):ar5.

Ken Hiltner. (2017). A nearly carbon-neutral conference model. last visited 2019-11-20.

Prpić, K. (2000). The publication productivity of young scientists: An empirical study. *Scientometrics*, 49(3):453–490.

Rowe, N. (2018). When you get what you want, but not what you need: The motivations, affordances and shortcomings of attending academic/scientific conferences. *International Journal of Research in Education and Science*, 4(2):714–729.

Proceedings, pages 402–406
Paris, UNESCO Headquarters, 5-6 December, 2019.

Indigenous Language Technologies & Language Reclamation in Canada

Nathan Thanyehténhas Brinklow[1], Patrick Littell[2], Delaney Lothian[3], Aidan Pine[2], Heather Souter[4]
Queen's University[1], National Research Council Canada[2], University of Alberta[3], Prairies to Woodlands Indigenous Language Revitalization Circle[4]
nathan.brinklow@queensu.ca, {patrick.littell, aidan.pine}@nrc-cnrc.gc.ca,
delaney.lothian@gmail.com, hsouter@gmail.com

Abstract

There is an incredible diversity of Indigenous languages spoken and signed in Canada, and Indigenous communities are committed to revitalizing and reclaiming them despite over a century of oppressive governmental policies against their use. Harnessing technology to support these efforts shows promise, but there are common pitfalls to avoid. We briefly describe the social and historical context surrounding Indigenous language technology development and implementation in Canada. We argue that the benefits of technology largely come from those which demonstrate an understanding of the relevant historical context, are tailored to community goals, and emphasize process over product.

Keywords: language revitalization, technology, language reclamation

Résumé

Il existe une grande diversité de langues autochtones parlées et signées au Canada et les communautés autochtones s'investissent dans la revitalization et la reprise de ces langues, malgré des centaines d'années de politiques oppressives contre leur usage. L'utilisation de la technologie est prometteuse, mais il y a aussi des embûches à éviter. Nous décrivons brièvement le contexte social et historique au sujet du développement des technologies langagières autochtones et de leur mise en œuvre au Canada. Nous soutenons que les bienfaits de la technologie se produisent largement lorsque le contexte historique est compris, lorsque les buts communautaires sont intégrés et lorsqu'on met l'accent sur le fait que la démarche est plus importante que le produit.

1. Introduction

Within only a few decades, digital language technology has become a widespread and popular way to help with learning or studying a language. Through greater access to language data and first language speakers, and with high demand from language learners, many strides have been made in creating new technology-based linguistic methods and language learning resources. However, what is far newer is the application of these resources and technologies to the Indigenous[1] languages of people from colonized countries all over the world. A large portion of conventional language resources and technologies are built with the goal of teaching language for reasons of tourism and employment. By contrast, Indigenous language technologies predominantly aim to further efforts of language documentation, revitalization, and reclamation. For this reason, as well as the unique linguistic properties of many Indigenous languages (Littell et. al., 2018), the application of current language technologies and methods to create new ones for Indigenous languages is not a straightforward process. In Canada, there has been a recent increase in funding for the creation of Indigenous language technologies. Though, increased opportunity demands increased critical discussion as to what these technologies can and should look like.

[1]Throughout this paper, the term 'Indigenous' will refer to the First Nations, Métis, and Inuit peoples of Canada, and as it pertains to their respective languages and cultures. We follow the Canadian Federal Translation Bureau's capitalization guidelines.

2. Background

Canada is home to approximately 70 unique Indigenous languages belonging to 10 distinct language families. Of these languages, approximately 57% have less than 500 active speakers (Rice, 2019). The decline of Indigenous language transmission in Canada can be attributed to many institutional and societal factors related to colonization, such as residential schools and the Indian Act. Canada's residential school system's explicit purpose was to assimilate Indigenous people into colonial culture by ensuring generations of Indigenous children did not learn their language and culture. These boarding schools were only fully closed "in 1996 after a duration of almost 150 years" (Griffith, 2017). This has profoundly affected the language transmission rates amongst Indigenous peoples in Canada, resulting in fewer and fewer new speakers (Truth and Reconciliation Commission of Canada, 2015).

This history has resulted in Indigenous peoples across Canada developing complex feelings towards learning their ancestral languages and cultures. Through intergenerational trauma from residential schools and other institutional forces, many Indigenous people still feel residual shame around speaking their language (Jenni, Anisman, McIvor, & Jacobs, 2017). Similarly, many

Indigenous people in Canada and around the world do not see learning their ancestral language as being valuable, with majority languages being seen as the language of success and social integration (UNESCO Ad Hoc Expert Group on Endangered Languages, 2003, p. 2). Despite these barriers, there are programs aimed at turning the tide of language loss; from immersion schools and language nests to drop-in classes and mentor-apprentice programs (Dunlop, Gessner, Herbert, & Parker, 2018). The choice Indigenous people all over the country are making to learn and teach their languages, in light of the explicit efforts to suppress them, is a decidedly political act - an act of anticolonial resistance (Pine & Turin, 2017; see also Urla, 2012 & Roche, 2020).

Digital technology is also being employed for Indigenous language revitalization, but its implementation is relatively nascent. Currently there are only a few groups (e.g. the National Research Council of Canada, the University of Alberta, the First Peoples' Cultural Council, the Computer Research Institute of Montréal, etc.) developing technology that focuses on creating building blocks for language work and everyday use, such as fonts and keyboards, spell-checkers, dictionaries, phrasebooks, and predictive text. However, there is relatively little text and speech data available for nearly all Indigenous languages in Canada and so these technologies typically must rely on rule-based approaches (as opposed to data-driven ones).

While there is substantial grammatical variation among Canadian Indigenous languages, they are almost all characterized by a high degree of morphological complexity and polysynthesis (i.e. words are composed of many smaller parts) (Rice, 2008). They are also all substantially different grammatically from the languages in which the foundation of the vast majority of language tools and methods were developed (e.g., English, French, Spanish, etc). This limits the scope of these technologies and leaves a gap between what is possible for Indigenous languages versus what currently exists for more prevalent languages, such as English (Littell et al., 2018).

3. Reframing 'Technology'

In contemporary speech, 'technology' and 'digital technology' are synonymous. In reality, digital technology is built on the shoulders of many other forms of technology. For example, one of the foundational technologies associated with computation is character encoding. Character encoding is how letters and other characters are numerically represented, and is an idea inherited from non-computational technologies like morse code and the telegraph. Given the ill-defined nature of the contemporary use of 'technology', we adopt a more functional definition. We define technology as a *force multiplier*: a tool, idea, object, or technique that allows people to accomplish their goals more quickly or by using fewer resources (see also Rice & Thieberger, 2018). Reframing the definition of technology from something tool-oriented to something goal-oriented allows technologists and users of technology to be more thoughtful and deliberate about the reasons to employ, or not to employ, a given technology.

When applying a technology to a language revitalization project in Canada, there are many goals to consider. While a primary goal of language revitalization projects is to encourage the use and learning of the language in question, it is rarely the only goal. In an address about the impact of the declaration of the UN International Year of Indigenous Languages, Dr. Lorna Williams stated "in the country that I come from, Indigenous languages were considered, and continue to be considered, of no value [...] our work has been to change this" (Williams, 2019). For many in Canada, like Dr. Williams, the goals of language revitalization transcend linguistic competence; they are connected to identity and community building, cultural resurgence, and broader social goals of self-determination. In recognition of this diversity of goals, some scholars and language advocates are using the term 'language reclamation' in place of revitalization (Leonard, 2012). In other words, the goals of language revitalization are only a subset of the goals of language reclamation.

4. Guiding Principles

In the context of Indigenous language reclamation, we assert that the process of developing language technology is as important as the product. Both the product and the process have the potential to empower or disempower language reclamation communities in equal measure (Alia, 2009, p.173). An example of this danger could be a potential technology that promotes language use but separates a language community from its data; essentially a tool that serves the goal of language revitalization while stifling the goal of language reclamation. This section discusses some of the guiding principles that can help ensure that technological development supports both language revitalization and the social and political goals of language reclamation.

4.1 Technology as a 'MacGuffin'

Technological solutions are frequently oversold in the media (e.g., 'New app is saving endangered language'), and expectations for new technology projects can be unrealistically high. Richard Grounds (2016) echoes this concern, asserting that, "[t]he very notion that these technological solutions somehow represent a kind of comprehensive and easy fix can itself become a problem that stands in the way of finding more effective directions for growing new fluent speakers. And this too often leads to diverting energy away from more effective paths for restoring the strength of our languages." To properly scope a given technology, it is necessary to have a firm understanding that it is *people* that revitalize a language and that technology can merely multiply their efforts. In other words, technology is the icing and not the cake.

It is routine for the 'inherent good' of digital technology to be assumed, regardless of how well it supports language community goals, and for its power and benefits to be exaggerated. In some cases, the technology itself serves as a sort of emblem of prestige much like the partial function of print dictionaries in the Pacific Northwest. While this is a beneficial aspect of technological development (Rice & Thieberger, 2018, p. 236; see also Ogilvie, 2011), new digital technology projects should strive to push past prestige to engage more directly with language communities' goals.

However, even functionally limited technology can concretely further language revitalization goals in its capacity to catalyze and motivate people. As technology development is inherently multi-disciplinary, it can bring together visual artists and craftspeople, musicians, gamers, and others who may not have previously thought of themselves as engaged in language reclamation. This bringing-together of people has sometimes achieved a greater goal than the comparatively minor goal the technology itself was intended to address. This has been seen anecdotally in the development of basic online phrasebooks and dictionaries (see Littell et. al. 2017). In many cases the pedagogical value of these tools for creating new speakers is rather limited; their primary function is as a reference tool. However, when young language learners understand how quickly linguistic data can be published online, it often catalyzes interest in creating more content. The creation of more content for an online dictionary, wordlist or phrasebook provides a focus for these young learners to build relationships and collaborate with elder generations. Strengthening intergenerational relationships is often the fundamental goal of language reclamation projects; a goal which can often surpass the impact of the dictionary or phrasebook itself.

We term such technologies *MacGuffins*: a literary term describing an object of perceived value that moves the plot forward, but which holds little intrinsic value. Many Indigenous language technologies are MacGuffins; building community connections and capacity by their development even if their stated goals are minor or only partially achieved. Put another way, the process of technology development can be valued alongside and even beyond the resulting technology. Realizing this benefit, however, requires planning for it by considering at the outset not just what the product is supposed to do, but who can benefit from their inclusion in the development process.

4.2 Indigenous User Experience

In order for technology to become a supportive stage for Indigenous language reclamation, those involved in the development must have an understanding of how Indigenous language learning differs from conventional language learning and how this will shape the Indigenous user experience. This includes everything from cultural implications of colonization to the appropriate and desired content of a given technology.

There are many ways in which shame may prevent an Indigenous person from learning their ancestral language (Jenni, Anisman, McIvor, & Jacobs, 2017). As mentioned previously, Canada subjected Indigenous people to forced assimilation for generations. It became unsafe and even illegal for some communities to practice their culture and speak their language (Henderson, 2018). Moreover, many people became isolated from their communities through Indian Act control of residency rights which affected, among others, women who married non-Indigenous men and individuals who attended university (Henderson, 2018). Indigenous language learners have cited this residual shame and fear of failure overall as a major obstacle. In this case, technology has a potential role to play in helping learners reach some level of fluency before interacting with speakers (Lothian, Akcayir, & Demmans Epp, 2019).

Another unique factor of Indigenous language technology is that Indigenous languages have historically been strictly oral and this requires consideration and accomodation in the development process. For example, some stories or names are not to be spoken of except by certain people or in certain seasons, and some cultural knowledge is expected to be received in person from elders and knowledge keepers. Accommodating this would require ensuring that the data used to develop content and linguistic models is appropriate and gathered by, or closely with, the relevant community.

4.3 Data Sovereignty & Open Source

Data sovereignty is an area of increasing concern for many Indigenous communities in Canada and internationally (Keegan, 2019). This discussion is rooted in the long history of exploitation by successive colonial governments, which now extends to the potential for exploitation and alienation of Indigenous data (Pool, 2016). In an era of widespread language reclamation activities, many communities identify their language data as a precious resource to be protected. Language technology projects must therefore recognize this reality and ensure that communities are able to protect themselves and their data from colonization in digital spaces (Dyson, Hendricks, & Grant, 2007).

With the staggering level of linguistic diversity in Canada, there is neither the time nor the resources to re-invent technologies each time they are applied to a new language. The language technology ecosystem that supports language revitalization must therefore be open source and well documented. This approach has the advantage of both reducing the level of investment in developing technology, and breaking down some of the barriers for collaboration and participation in the development process. Proprietary solutions, and solutions

which are undocumented to the point of being inaccessible, reinforce the power of a small group of experts at the expense of the larger language revitalization community.

The benefit of open sourcing tools is clear in the low-resource context in which language reclamation is taking place. However, given concerns about data sovereignty as it relates to language reclamation, 'open source' requires a more nuanced application in the Indigenous context, especially at the interface between 'tool' and 'data'. This tension is captured by the Kaitiakitanga License (Te Hiku Media, 2018), which is rooted in Maori community values loosely translated as 'guardianship.' The Kaitiakitanga license model was developed by an Indigenous organization in response to their concern that "by simply open sourcing our data and knowledge, we further allow ourselves to be colonised digitally in the modern world." (Te Hiku Media, 2018). In the Canadian context, the discussion is ongoing and is indicative of the distinction between language revitalization and language reclamation.

5. Conclusion

The prioritization of Indigenous community needs and goals has clear consequences for partners from government, academia, and industry engaged in the development of language technology, including awareness of community cultural values and protocols. The autonomy and agency of Indigenous communities must be recognized in all stages of development as the community decides how to develop and engage with language technologies. This approach allows communities to contribute in meaningful ways to linguistic and cultural continuity through technology as part of language reclamation.

There is clear potential for useful and well-designed Indigenous language technologies in Canada to support language revitalization. However, in the context of language reclamation, responsible technology development must engage with these matters directly. The conversation about Indigenous language technologies is not just about building the right tools, it is about building the tools in the right way—a way that recognizes and affirms the broader social and political goals of language reclamation.

6. Acknowledgements

We would like to acknowledge and thank the organizers of LT4ALL for taking the initiative on creating such an important international forum. We would also like to thank Riplea Lothian, Nicki Benson, and Zoë Cilliers for editorial comments and suggestions.

References

Alia, V., & Berghahn Books. (2009). The new media nation: Indigenous peoples and global communication. New York: Berghahn Books.

Dunlop, B., Gessner, S., Herbert, T., & Parker, A. (2018). Report on the status of B.C. First Nations languages. Retrieved from http://www.fpcc.ca/files/PDF/2010-report-on-the-status-of-bc-first-nations-languages.pdf

Dyson, L. E., Hendriks, M., & Grant, S. (Eds.). (2007). Information technology and Indigenous peoples. Hershey, PA: Information Science Publishing.

Grounds, Richard A. 2016. Indigenous Perspectives and Language Habitats. Paper presented to the *International Expert Group Meeting on Indigenous Languages: Preservation and Revitalization*. United Nations, Department of Economic and Social Affairs, New York, January 19–21. http://www.un.org/esa/socdev/unpfii/documents/2016/egm/ Paper_Grounds2.pdf.

Griffith, J. (2017) Of linguicide and resistance: Children and English instruction in nineteenth-century Indian boarding schools in Canada. *Paedagogica Historica*, 53:6, 763-782, DOI: 10.1080/00309230.2017.1293700

Henderson, W. (2018). Indian Act In *The Canadian Encyclopedia*. Retrieved from https://www.thecanadianencyclopedia.ca/en/article/indian-act

Jenni, B., Anisman, A., McIvor, O., & Jacobs, P. (2017). An exploration of the effects of mentor-apprentice programs on mentors' and apprentices' wellbeing. *International Journal of Indigenous Health*, 12(2), 25. https://doi.org/10.18357/ijih122201717783

Keegan, Te Taka. (2019, June 25). Issues with Māori sovereignty over Māori language data [Video File]. Retrieved from http://video.web.gov.bc.ca/public/fpcc/letlanguageslive.html

Leonard, W. Y. (2012). Framing language reclamation programmes for everybody's empowerment. *Gender and Language,* 6(2), 339–367.

Littell, P., Kazantseva, A., Kuhn, R., Pine, A., Arppe, A., Cox, C., & Junker, M. (2018). Indigenous language technologies in Canada: Assessment, challenges, and successes. In J. L. Klavans (Ed.), *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 2620–2632). Retrieved from https://www.aclweb.org/anthology/W18-1921

Littell, P., Pine, A., & Davis, H. (2017). Waldayu and Waldayu Mobile: Modern digital dictionary interfaces for endangered languages. In Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages (pp. 141-150).

Lothian, D., Akcayir, G., & Demmans Epp, C. (2019). Accommodating Indigenous People When Using Technology to Learn Their Ancestral Language. *International Workshop on Supporting Lifelong Learning at the 20th International Conference on Artificial Intelligence in Education (AIED)* (Vol. 2395

pp. 16-22), Chicago, Illinois, USA. CEUR-Workshop Proceedings.

McIvor, O., & Anisman, A. (2018). Keeping our languages alive: Strategies for Indigenous language revitalization and maintenance. In Y. Watanabe (Ed.), *Handbook of Cultural Security* (pp. 90–109). https://doi.org/10.4337/9781786437747.00011

Ogilvie, S. (2011). Linguistics, lexicography, and the revitalization of endangered languages. *International Journal of Lexicography*, 24(4), 389-404. doi:10.1093/ijl/ecr019

Pine, A., & Turin, M. (2017). Language Revitalization. Oxford Research Encyclopedia of Linguistics. Retrieved from, https://oxfordre.com/linguistics/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-8

Pool, I. (2016). Colonialism's and postcolonialism's fellow traveller: The collection, use and misuse of data on Indigenous people. In T. Kukutai & J. Taylor (Eds.), *Indigenous Data Sovereignty* (pp. 57–76). Retrieved from https://www.jstor.org/stable/j.ctt1q1crgf.11

Rice, K. (2019). Indigenous languages in Canada. In *The Canadian Encyclopedia*. Available at https://www.thecanadianencyclopedia.ca/en/article/aboriginal-people-languages

Rice, K., & Thieberger, N. (2018). Tools and technology for language documentation and revitalization. In K. L. Rehg & L. Campbell (Eds.), *The Oxford Handbook of Endangered Languages*. https://doi.org/10.1093/oxfordhb/9780190610029.013.13

Roche, G. (2020). Language revitalization and radical politics. Language on the Move. https://www.languageonthemove.com/language-revitalization-and-radical-politics/

Te Hiku Media. (2018). Kaitiakitanga License. Retrieved from https://github.com/TeHikuMedia/Kaitiakitanga-License/blob/master/LICENSE.md

Truth and Reconciliation Commission of Canada. (2015). Honouring the truth, reconciling for the future. Retrieved from http://www.trc.ca/assets/pdf/Honouring_the_Truth_Reconciling_for_the_Future_July_23_2015.pdf

UNESCO Ad Hoc Expert Group on Endangered Languages. (2003). Language vitality and endangerment. Retrieved from http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/CLT/pdf/Language_vitality_and_endangerment_EN.pdf

Urla, J. (2012). Reclaiming Basque: Language, nation, and cultural activism. University of Nevada Press.

Williams, L. (2019, Dec 6). "Technology: Is it a help or a hindrance?" [Video File]. Retrieved from http://webcast.unesco.org/live/vod/2019/ci/20190612_ci_room-02/en/

Proceedings, pages 407–436
Paris, UNESCO Headquarters, 5-6 December, 2019.

LT4All 2019 Oceanian Languages Poster Session

P.1.1: Building capacity for community-led documentation in Erakor, Vanuatu
Authors: Ana Krajinovic, Rosey Billington, Lionel Emil, Gray Kaltaƥau and Nick Thieberger
Country: Germany

Abstract: We discuss a collaboration between community members and visiting researchers in Erakor, Vanuatu aiming to build the capacity of community-based researchers to undertake language and cultural documentation projects. We focus on the outcomes and benefits of the community-led project in Erakor after initial training, which include: a) long-term documentation of linguistic and cultural practices calibrated towards community's needs (using the PARADISEC repository for ongoing access), and b) collections of large quantities of data of good phonetic quality, which, besides being readily available for research, have a great potential for training and testing language technologies, such as automatic speech recognition.

Native language: Komam utilusus teflan naɱer nig natkon go naɱer weswes ni nlaun nakon nen ruto saof natkon, teflan rufaitau go rutafnau weswes ni natkon raki teflan ruktɱer wesweski nawesien ni namtrirsokwen go nakraksokien ni nafsan go suƥ. Komam ule toop pak nua nawesien ni teflan nataɱol ni natkon ɱas itɱen weswes eɱrom ni Erakor, go ntakun ni tete nfaitauen go nafregnrogwen toklos, a) nawesien ni nakraksokien ni nafsan go suƥ raki naɱitwen nig nfanu, b) nakraksokien ni data kelaap go tenen misleo knen iwi, nen rukta tae ler nametmatuan knen ɱas mau me nen ruktae pregi rupi teni nafregnrogwen ni nanrogwen ni nafeswen.

P.1.2: PARADISEC (Pacific and Regional Archive for Digital Sources in Endangered Cultures)
Authors: Amanda Harris and Nick Thieberger
Country: Australia

Abstract: Language archives play an important role in keeping records of the world's languages safe. Accessible audio recordings held in archives can be used by speakers of small and endangered languages, and their communities, and provide a base for further research and documentation. There is an urgent need for historical analog tape recordings to be located and digitised, as they will soon be unplayable. PARADISEC holds records in 1228 languages. We run training for language documentation and are developing technologies to localise access to language records. A concerted effort is needed to support language archives and sustain language diversity.

Resume: Wanpela kain ples olsem akaiv i save lukautim ol rekod or pepa bilong ol kain kain tok ples bai stap gut long bihain taim. Planti ol liklik tok ples ol klostu dai nau. Tok ples bilong ol manmeri na komuniti mas usim akaiv long helpim wok bilong painim aut moa na raitim ol pepa bilong ol tok ples. PARADISEC i gat 1228 tok ples. Mipela painim ol rikoding long taim tumbuna we ol tok ples i stap long keset tep bilong dijitaisim nau o sapos nogat bai ol bagarap. Taim nau long sapotim ol tok ples akaiv long lukautim planti kain kain tok ples.

P.1.3: Bloom Books
Authors: Paul Nelson
Country: United States

Abstract: Bloom is a tool/system that facilitates the creation of literacy materials in any language. Beginning with shell books, a user can simply translate the stories into many languages to build libraries of books. Bloom books can be published as PDFs that are printed, ePubs, Bloom Reader apps (to be read on Android devices) and web sites.

Bloom Enterprise provides extra functionality to create books for the deaf, comprehension questions to ascertain if the readers understand, branding for organizations who sponsor literacy programs, and a dashboard to understand what books are being read, for how long, and how comprehension is trending.

P.1.4: Creating a Synthetic Te Reo Māori Voice
Authors: Isabella Shields, Catherine Watson, Peter Keegan, Rebekah Berriman and Jesin James
Country: New Zealand

Abstract: We have made a synthetic male te reo Māori voice, which runs on MaryTTS and synthesises speech from any Māori text. The voice was created from recordings of 1030 sentences, chosen to ensure full diphone coverage, from the story Ngā Mahi a Ngā Tūpuna. The recordings were made in a soundproof booth and condensed to two hours of continuous speech. Phonetic labelling was first determined automatically using a model obtained from Montreal Forced Aligner. The labelled data, along with a 10,000-word Māori lexicon with phonetic transcription and stress mark up, is passed into MaryTTS to create a synthetic voice.

Native language: Kua hangaia e mātou he reo rorohiko, he reo tāne. Ka noho tēnei reo rorohiko ki roto i a te pūmanawa MaryTTS. Ka whakaurua e tangata he tuhinga, ā, ka hurihia he reo, arā he kōnae tangi. Kua hopukia tēnei ki roto i tētahi wharau karo tangi. I ahu mai ngā kōrero i te pukapuka Ngā Mahi a Ngā Tupuna, ā, kua wehewehe ngā kōrero kia rerenga kōrero poto (1030). Ka āta whiriwhiria ētahi anō kia mau te katoa o ngā tangi oro o te reo Māori. Kia hangai atu ngā oro tuhi ki ngā oro tangi, ka whakamahia te Montreal Forced Aligner, nā te tangata anō i atā tirotiro te tika o ngā hononga. Ka tukua ēnei raraunga ki a MaryTTS, ka tukua hoki he papakupu whakaahua (10,000 + ngā kupu).

P.2.1: Poio - Open Source Technology for Language Diversity
Authors: Peter Bouda
Country: Portugal

Abstract: The Poio project publishes open source tools to support under-resourced languages on computers and mobile devices. Our main product is the Poio Text Prediction, a corpus-based text input support system that simplifies the way people enter text in any language. We believe that technology should assist the renewal of local languages and cultures by allowing people to actively teach, learn, extend, and spread their language in their community. Our aim is to give people the ability to use their mother tongue in everyday electronic communication, no matter where they are or what language they speak.

Native language: Das Projekt Poio entwickelt Open-Source-Lösungen um Sprachen auf Computern und mobilen Geräten zu unterstützen für die es nur wenige Daten gibt. Unser Hauptprodukt ist Poio Text Prediction, ein korpus-basiertes Eingabeunterstützungssystem, das jedem die Eingabe jeder beliebigen Sprache erleichtert. Wir glauben, dass Technologie die Erneuerung lokaler Sprachen und Kulturen unterstützen sollte, indem sie Menschen erlaubt ihre Sprache in ihrer Gemeinschaft zu lehren, zu lernen, zu erweitern und zu verbreiten. Unser Ziel ist es allen Menschen die Möglichkeit zu geben ihre Sprache in alltäglicher, elektronischer Kommunikation zu verwenden egal wo sie sich befinden oder welche Sprache sie sprechen.

P.2.2: Towards a Global Lexicographic Infrastructure
Authors: Simon Krek, Thierry Declerck, John Philip McCrae and Tanja Wissik
Country: Germany

Abstract: The poster describes the European Lexicography Infrastructure (ELEXIS), showing how it includes, integrates and cross-links also non-European dictionaries, and how anyone can contribute, either with data, scientific exchange or with an institutional cooperation, which can be implemented in the form of Observer status.
ELEXIS provides for tools and services to develop new dictionary databases or process and enrich existing ones. It also provides for grants for short scientific visits for helping the visitor to get acquainted with the ELEXIS technologies. With its linking strategies, ELEXIS ensures that lexical data of each language is getting a high visibility and accessibility.

Native language: L'affiche décrit l'infrastructure lexicographique européenne (ELEXIS), en montrant comment elle inclut, intègre et interconnecte les dictionnaires (également non européens), et comment chacun peut contribuer, soit par des données, des échanges scientifiques ou une coopération institutionnelle, qui peut être mise en œuvre sous la forme du statut d'observateur.
ELEXIS fournit des outils et des services pour développer de nouvelles bases de données de dictionnaires ou pour traiter et enrichir des bases de données existantes. Il prévoit également des bourses pour de courtes visites scientifiques afin d'aider le visiteur à se familiariser avec les technologies ELEXIS. Grâce à ses stratégies de mise en relation, ELEXIS assure une grande visibilité et accessibilité aux données lexicales de chaque langue.

P.2.3: Tooling up a less-resourced language with NLP : the example of Corsican and the "Banque de Données Langue Corse" (BDLC, Corsican Language Database)
Authors: Laurent Kevers, Stella Retali-Medori, Florian Guéniot and A. Ghjacumina Tognotti
Country: France

Abstract: The current situation regarding the existence of Natural Language Processing (NLP) resources and tools for Corsican reveals their virtual non-existence. Our inventory contains only a few rare digital resources, lexical or corpus databases, requiring adaptation work. Our objective is to use the BDLC project to improve the availability of resources and tools for the Corsican language. We have defined a roadmap setting out the actions to be undertaken: collection of corpora and setting up of a consultation interface (concordancer), language detection tool, electronic dictionary and part-of-speech tagger. The first achievements are already available.

Native language: L'état des lieux concernant les ressources et outils de Traitement Automatique du Langage (TAL) pour le corse révèle leur quasi inexistence. Notre inventaire ne contient que quelques rares ressources digitales, bases de données lexicales ou corpus, nécessitant un travail d'adaptation. Nous nous appuyons sur le projet BDLC pour faire avancer l'outillage de la langue corse. Nous avons défini une feuille de route reprenant les actions à entreprendre : collecte de corpus et mise en place d'une interface de consultation (concordancier), outil de détection de langue, dictionnaire électronique et outil d'annotation en parties du discours. Les premières réalisations sont déjà disponibles.

P.2.4: Language Technology Program for Icelandic
Authors: Anna Nikulásdóttir
Country: Iceland

Abstract: On the 1st of October 2019, work on a five-year Project Plan for Icelandic LT started. The project aims at the development of LT resources and infrastructure software, including speech technologies, machine translation and spell and grammar checking systems. It emphasizes cooperation between academia and industries, with the aim of open and usable software and resources for LT-products being delivered at the end of the program. The self-owned foundation Almannarómur conducts the program on behalf of the Icelandic Government. The research and development work is carried out by teams built across a consortium of nine universities, institutions, and private companies.

Native language: Þann 1. Október 2019 hófst vinna við fimm ára Verkáætlun í máltækni fyrir íslensku. Markmið áætlunarinnar er að þróa málföng og innviði fyrir máltækni, þar sem áhersla verður lögð á taltækni, vélþýðingar og málrýni. Lögð er áhersla á samstarf milli háskólasamfélagsins og fyrirtækja með það að markmiði að áætlunin skili opnum og nothæfum málföngum og hugbúnaði til notkunar í hugbúnaði sem þarfnast máltækni. Sjálfseignarstofnunin Almannarómur stýrir áætluninni fyrir hönd Ríkisstjórnar Íslands. Rannsóknar- og þróunarvinna er unnin af teymum sem mynduð eru í samstarfi níu háskóla, stofnanna og einkafyrirtækja.

P.2.5: A speaking atlas of indigenous languages of France and its Overseas
Authors: Philippe Boula de Mareüil, Gilles Adda, Albert Rilliard and Frédéric Vernier
Country: France

Abstract: The objective is to valorise the linguistic diversity of France through field recordings, a computer-based visualisation of dialectal areas and orthographic transcripts (which represent an object of research in itself). We describe here a website (https://atlas.limsi.fr) presenting interactive maps of Metropolitan France and its Overseas, from which the Aesop fable "The Northwind and the Sun" can be listened to and read in over 300 versions, in regional languages. There is thus both a scientific dimension and a heritage dimension in this work, since a number of regional or minority languages are in a critical situation.

Native language: L'objectif est de montrer et de valoriser la diversité linguistique de la France à travers des enregistrements recueillis sur le terrain, une réalisation informatique (qui permet de visualiser les aires dialectales) et un travail de transcription orthographique. Nous décrivons ici un site web (https://atlas.limsi.fr) présentant des cartes interactives de France hexagonale et des Outre-mer, à partir desquelles la fable d'Ésope «La bise et le soleil » peut être écoutée et lue dans plus de 300 versions, en langues régionales. Il y a ainsi une dimension à la fois scientifique et patrimoniale à ce travail, dans la mesure où un certain nombre de langues régionales ou minoritaires sont en situation critique.

P.2.6: Software and Linguistic Resources for the Tatar language preservation and development: Regional Experience
Authors: Dzhavdet Suleymanov, Aidar Khusainov and Rinat Gilmullin
Country: Russian Federation

Abstract: The poster contains information about the most significant program developments and linguistic resources of the Institute of Applied Semiotics of Tatarstan Academy of Sciences, aimed at supporting the Tatar language in information technologies: the National Corpus of the Tatar language "Tugan tel", the Russian-Tatar machine translation system Tatsoft, Tatar speech synthesis and recognition systems and mobile applications. The main results achieved within of the State program for the preservation, study and development of the state languages of the Republic of Tatarstan and other languages in the Republic of Tatarstan.

Native language: Постерда Татарстан Республикасы Гамәли семиотика Институты тарафыннан башкарылган иң әһәмиятле программа һәм лингвистик ресурслар турында мәгълумат күрсәтелә. Алар арасында – “Туган тел” татар гомумтел корпусы, “Татсофт” русча-татарча тәрҗемәче программа, Татар сөйләмен тавышландыру һәм текстка күчерү системалары, Смартфоннар өчен клавиатура һәм сүзлекләр кушымталары.
Башкарылган эшләрнең төп нәтиҗәләре Татарстан Республикасы дәүләт телләрен һәм Татарстан Республикасындагы башка телләрне саклау, өйрәнү һәм үстерү буенча Татарстан Республикасы дәүләт программасын гамәлгә ашыру кысаларында гамәлгә ашырылды.

P.2.7: Contribution to the Universal Dependencies Treebank of Non-Standard Romanian Texts
Authors: Victoria Bobicev, Catalina Mărănduc, Tudor Bumbu, Ludmila Malahov, Alexandru Colesnicov and Svetlana Cojocaru
Country: Republic of Moldova

Abstract: Cultural heritage preservation is the one non-transferable duty of any given ethnic or social entity, for it is the essence that defines and identifies each one of them among others. In the specific case of the preservation of culturally significant works of writing, this task includes not only digitizing old books to prevent their loss but also optical character recognition, transliteration of old texts and their annotation. We report our latest contribution to the development and enrichment of a universal dependencies (UD) treebank which contains old texts, regional folklore and other non-standard texts from Moldova and Romania.

Native language: Păstrarea patrimoniului cultural este datoria netransmisibilă a oricărei entităţi etnice sau sociale, deoarece este esenţa care o defineşte şi identifică. În cazul specific al conservării operelor literare semnificative din punct de vedere cultural, această sarcină include nu numai digitalizarea cărţilor vechi pentru a preveni pierderea lor, dar şi recunoaşterea optică a caracterelor, transliterarea textelor vechi şi adnotarea lor. Raportăm contribuţia noastră recentă la dezvoltarea Treebank-ului de dependenţe universale (UD) care conţine texte vechi, folclor regional şi alte texte non-standard din Moldova şi România.

P.2.8: Inquiring about digital use and usability of minority languages: the approach of the Digital Language Diversity Project
Authors: Claudia Soria and Cor van der Meer
Country: Italy

Abstract: We present the results of the Digital Language Diversity Project survey about the digital behaviour, desires, and expectations of minority language speakers. The survey is designed around three conceptual blocks: the digital capacity of the language, its digital opportunities, and speakers' attitudes towards digital use of the language. We believe that the DLDP initiative has the potential to be extended to other languages and deserves to be considered by the community at large as a useful tool for digital language planning. See also http://wp.dldp.eu/reports-on-digital-language-diversity-in-europe/.

Native language: Presentiamo i risultati del sondaggio del Digital Language Diversity Project sul comportamento digitale, i desideri e le aspettative dei parlanti di alcune lingue minoritarie. Il sondaggio è strutturato attorno a tre blocchi concettuali: la capacità digitale del linguaggio, le sue opportunità digitali e gli atteggiamenti dei parlanti nei confronti dell'uso digitale del linguaggio. Riteniamo che l'iniziativa DLDP abbia il potenziale per essere estesa ad altre lingue e meriti di essere considerata dalla comunità in generale come uno strumento utile per la pianificazione linguistica digitale. Vedi anche http://wp.dldp.eu/reports-on-digital-language-diversity-in-europe/.

P.2.9: Language Technologies for Istro-Romanian

Authors: Patricia Serbac

Country: Romania

Abstract: Istro-Romanian is a minority language spoken in the peninsula of Istria, Croatia. It is a severely endangered language, with about 100 speakers left. The last speakers originate from woo Istrian villages, but they have spread now in the rest of the peninsula, in Croatia, in the neighbouring countries and in the world. Due to emmigration, there are Istro-Romanians also in the US and in Australia. The poster presents the own audio and video corpus of Istro-Romanian, the webpage of the community, mail communication etc. The new language technologies could help keeping these speakers together and making them one community.

Native language: Istro-româna e o limbă minoritară vorbită în peninsula Istria, din Croația. E o limbă în pericol grav, având în jur de 100 de vorbitori. Ultimii vorbitori provin din două sate din Istria, dar ei s-au răspândit în restul peninsulei, în Croația, în țările învecinate și în lume. Din cauza emigrării, există istro-români și în SUA și în Australia. Posterul prezintă propriul corpus audio and video al istro-românei, pagina de internet a comunității, comunicarea pe mail etc. Noile tehnologii despre limbă îi pot ajuta pe acești vorbitori să se adune într-o singură comunitate.

P.2.10: Innovative CALL Solutions and The Sustainability of "Nano" Languages in the West-Nordic Arctic Region

Authors: Birna Arnbjörnsdóttir and Auður Hauksdóttir

Country: Iceland

Abstract: This poster describes five innovative CALL solutions that support multilingualism and „nano“ languages in the West-Nordic Region. The non-language specific platforms are developed at the Vigdís Finnbogadóttir Institute at the University of Iceland and include open curated language courses, www.icelandiconline.com, and www.faroese.fa; and tools that enhance oral fluency and communicative skills in Danish as a second and third language in the West-Nordic region. The tools include: www.talboblen.hi.is, that focuses on oral language skills, www.talerum.is , an interactive game based program that encourages interaction, and www.frasar.net a resource that teaches the pragmatics of phrases.

Native language: Veggspjaldið lýsir fimm nýstárlegum tæknilausnum sem styðja fjöltyngi og heimamál á Vestnorræna málsvæðinu. Kerfin eru þróuð við Stofnun Vigdísar Finnbogadóttur við Háskóla Íslands og eru óháð tungumálum. Lýst verður opnum stýrðum tungumálanámskeiðum í íslensku og færeysku, www.icelandiconline.com og www.faroese.fa; auk tæknilausna sem styðja við fjöltyngi og auka munnlega færni í dönsku sem öðru og þriðja máli. Þau kerfi eru www.talboblen.hi.is sem eflir talmál og framburð, www.talerum.is sem er gagnvirkur tölvuleikur sem hvetur til ritunar og talmálsnotkunar og www.frasar.net sem eflir viðeigandi notkun frasa á íslensku og dönsku.

P.2.11: Indigenous/Minority Language Keyboard and Spell Checking Support, for Desktop and Mobile Operating Systems

Authors: Brendan Molloy

Country: Sweden

Abstract: Can you spell out your own name with your current keyboard? Members of several indigenous language communities cannot, ignored by the technological revolution in how we communicate.

Divvun, in collaboration with The Techno Creatives, provides keyboards for mobile and desktop environments, with spell checking integration for Sámi languages and other languages with very complex morphology. We even include Mansi support for this conference!

Our keyboard integration for iOS is so robust you can test our minority language support through QR codes provided on this poster.

Experience a level of minority language support Google and Apple have failed to provide.

Native language: Kan du skriva namnet ditt med tastaturet du har framfor deg? Folk frå mange urfolkssamfunn kan ikkje det, ignorerte som dei er av den tekniske revolusjonen i måten vi kommuniserer på.

Divvun-gruppa, i lag med The Techno Creatives, tilbyr tastatur for mobiltelefonar og datamaskiner, med innebygd stavekontroll for samiske og andre språk med svært kompleks morfologi. Vi har til og med lagt til støtte for mansisk for denne konferansen!

Tastaturingegreringa vår for iOS er so robust at du kan testa støtta for minoritetsspråk med QR-kodane som finst på denne plakaten.

Opplev eit nivå på minoritetsspråkstøtta som Google og Apple ikkje har lyktes med å levera.

2nd Native language: [illegible]

Divvun-[illegible], [illegible] The Techno Creatives, [illegible]. [illegible] iOS [illegible] QR-[illegible].

[illegible] Google [illegible] Apple [illegible].

P.2.12: MultiTAL : an online platform to list NLP tools for under-resourced languages

Authors: Damien Nouvel, Driss Sadoun and Mathieu Valette

Country: France

Abstract: Diversity and variety of human languages raises indisputable difficulties for processing textual data. Regarding under-resourced languages, many softwares have been implemented, but many are poorly referenced and documented. The ERTIM (INALCO) lab has published in 2016 a website (http://multital.inalco.fr) that addresses this issue. Our website lists tools available for languages. For each software, the database provides information concerning : NLP tasks, implemented method, OS compatibility, among others. We do not pretend to be exhaustive, but people populating the database are speakers of concerned languages, they downloaded and tested softwares, and provided technical information for their installation and use.

Native language: La diversité et la variété des langues humaines donne d'incontestables difficultés pour le traitement de données textuelles. Concernant les langages peu dotés, de nombreux logiciels ont été implémentés, mais beaucoup restent peu référencés et mal documentés. L'équipe ERTIM a mis en ligne en 2016 un site (http://multital.inalco.fr) qui réalise ce travail. En faisant la liste des outils par langage, cette base de données apporte des informations sur leur utilisation. Nous ne prétendons pas être exhaustifs, mais les personnes remplissant la base étaient locuteurs des langues concernées, elles ont téléchargé et testé les outils, et ont renseignés des informations sur leur installation et leur utilisation.

P.2.13: Automatic Recognition of mixed Ukrainian-Russian Speech

Authors: Valeriy Pylypenko and Tetyana Lyudovyk

Country: Ukraine

Abstract: This work presents an approach to recognition of conversational speech with code-switching which is widespread in Ukraine now. Both inter-sentential and intra-sentential Ukrainian/Russian code-switching is handled. The approach takes into account closely related Russian and Ukrainian phonetic systems. A cross-lingual ASR system is developed. The acoustic model and pronunciation lexicon are based on Ukrainian phone set. Experiments with different types of code-switching speech (Parliamentary, TV broadcast) were conducted and results are presented. The approach is suitable especially in cases of intra-sentential code-switching where language identification is problematic.

Native language: Ця робота представляє підхід до розпізнавання усного мовлення з переключенням між українською та російською мовами, яке зараз поширене в Україні. Обробляється як міжфразове перемикання, так і всередині фраз. Підхід враховує особливості фонетичних систем тісно пов'язаних російської та української мов. Розроблена багатомовна система автоматичного розпізнавання мовлення. Акустична модель та лексика вимови базуються на українській множині фонем. Представлені результати розпізнавання мовлення з переключенням мови з декількох джерел (Парламентська, ТВ трансляція). Підхід особливо корисний у випадках перемикання мови всередині фраз, де ідентифікація мови є проблематичною.

P.2.14: Apertium: a free/open-source platform for machine translation and basic language technology

Authors: Mikel L. Forcada and Francis Tyers

Country: Spain

Abstract: Apertium is a free/open-source platform for rule-based machine translation and basic language technology. Since 2005, Apertium provides a free/open-source, modular, language-independent machine translation engine, free/open-source linguistic data for a variety of languages and language pairs, with emphasis on less-resourced languages, and free/open-source tools to manage language data, learn rules, and build machine translation engines. The Apertium pipeline contains monolingual modules useful in other human-language technology tasks. The license chosen, the GPL, avoids private appropriation and encourages giving improvements back to the project, creating a community. Apertium is an active research and business platform, and provides a series of stand-alone products.

Native language: Apertium és una plataforma lliure/de codi obert per a la traducció automàtica basada en regles i per a tecnologies bàsiques de la llengua. Des del 2005, Apertium proporciona un motor de traducció automàtica lliure / de codi obert, modular, independent de la llengua, dades lingüístiques lliures / de codi obert per a diversos idiomes i parells d'idiomes, amb èmfasi en llengües amb menys recursos i eines lliures/de codi obert per gestionar dades lingüístiques, aprendre regles i crear motors de traducció automàtica. El 'pipeline' d'Apertium conté mòduls monolingües útils en altres tasques de tecnologia del llenguatge humà. La llicència escollida, la GPL, evita l'apropiació privada i incita a retornar millores al projecte, creant una comunitat. Apertium és una plataforma activa de recerca i negocis i ofereix una sèrie de productes per a usuaris finals.

P.2.15: REDISCOVERING PAST NARRATIONS: THE ORAL HISTORY OF THE ROMANIAN LANGUAGE PRESERVED WITHIN THE NATIONAL PHONOGRAMIC ARCHIVE

Authors: Oana Niculescu, Maria Marin and Daniela Răuţu

Country: Romania

Abstract: Archive. A monumental work, AFLR is the richest, most inclusive and diversified collection of ethno-linguistic recordings in Romania. Only a third of the data have been processed so far and there is a pressing need to digitize the remaining tape recordings. Through the preservation of the archive we can gain access to both individual and collective memories, aiding to a better understanding of our cultural heritage and, at the same time, restoring missing or forgotten pieces of Europe's oral history.

Native language: În această prezentare ne propunem să atragem atenţia asupra necesităţii conservării şi digitalizării Arhivei fonogramice a limbii române. În momentul de faţă, AFLR este cea mai bogată şi cuprinzătoare colecţie de texte dialectale din România. Cu toate acestea, doar o treime din material a fost digitalizat, existând riscul ca benzile rămase să se deterioreze, ducând la pierderea înregistrărilor. Protejarea arhivei AFLR contribuie, pe de o parte, la recuperarea naraţiunilor individuale şi colective, iar, pe de altă parte, la înţelegerea şi valorificarea patrimoniului cultural, respectiv redobândirea unor elemente pierdute sau uitate din istoria orală a Europei.

P.2.16: Language technology for indigenous languages: Achievements and challenges
Authors: Sjur Moshagen, Lene Antonsen and Trond Trosterud
Country: Norway

Abstract: Fifteen years of indigenous language technology development by UiT/Sámi Parliament has resulted in spelling and grammar checkers, desktop/mobile keyboards, morphological analysers, MT, speech synthesis, language learning tools and intelligent electronic dictionaries.

This was facilitated by an open source language independent infrastructure, targeted at languages with rich and complex grammar, with integration for host operating systems and apps.

The current primary challenge is integration with closed platforms where we cannot currently support user needs.

Our proposed solution is a “Manifesto for Open Language Technology”, where APIs, localisations and source code are open, while ensuring community intellectual property custodianship, engagement and commitment.

Native language: UiT/Sámedikki 15 jagi eamiálbmot giellateknologiija barggu bohtosat leat sátne- ja grammatihkkadivvunprográmmat, boallobeavdi dihtorii ja mobiltelefovdnii, morfologalaš analysáhtorat, dihtorjorgaleapmi, hállansyntesa, giellaoahppanreaiddut ja intelligeanta digitála sátnegirjjit.

Dát lea huksejuvvon rabas gáldokoda infrastruktuvrras, mii lea heivehuvvon gielaide main lea rikkes ja kompleaksa grammatihkka – infrastruktuvra mii siskkilda geavahanlavttaid ja applikašuvnnaid.

Dál váldohástalus lea integreret prográmmaid giddejuvvon geavahanvuogádagaide, maid siste mii dál eat beasa doarjut geavaheddjiid dárbbsuid.

Min evttohus lea "Rabas giellateknologiija manifesta", mas API:t, lokaliseren ja gáldokoda leat rabas, muhto seammás giellaservodagat galget hálddašit gáldokoda intellektuealla rivttiid.

P.2.17: Using technology to empower Indigenous knowledge sharing
Authors: Lorna Williams, Tracey Herbert and Daniel Yona
Country: Canada

Abstract: B.C. is an Indigenous language hot spot with 34 languages. In response to threats to the vitality of these languages, communities in B.C. have adopted collaborative approaches to language revitalization and technology. FirstVoices is an example of a collaborative, community-led language revitalization platform where communities manage, curate and control their data. The platform encourages youth to connect with fluent elders, in order to share their knowledge in a digital space. This poster presents realities, challenges, and opportunities of language revitalization and the ripple effect technology, such as keyboards or shared platforms, can have on access to language.

P.2.18: European Language Monitor by EFNIL
Authors: Sabine Kirchmeier
Country: Denmark

Abstract: The poster presents The European Language Monitor (ELM) - a key project of EFNIL, the European Federation of National Institutions for Language. ELM is an online database containing data on language legislation and language planning in Europe. The user can browse and compare language legislation, information on minority languages, and provisions for language use in the educational systems, in business, in the media and for language technology.
ELM contains comments, quotes, links and translations of national legislation wherever possible. The data for ELM are collected every 4 years. The current version, ELM 4, is based on data collected in 2017-2018.

Native language: Denne poster præsenterer The European Language Monitor (ELM) – et nøgleprojekt for EFNIL, den europæiske sammenslutning af officielle sproginstitutioner. ELM er en online database som indeholder data om sproglovgivning og sprogplanlægning i Europa. Brugeren kan søge og sammenligne sproglovgivning, information om minoritetssprog og regulering af sprogbrugen i uddannelsessystemet, i erhvervslivet, i medierne og inden for sprogteknologi. ELM indeholder kommentarer, citater, links og oversættelser af national lovgivning hvor det er muligt. Data til ELM indsamles hvert 4. år. Den nuværende version, ELM 4, er baseret på data som er indsamlet I 2018-2018.

P.2.19: Preserving Endangered European Cultural Heritage and Languages Through Translated Literary Texts
Authors: Amel Fraisse, Ronald Jenn, Shelley Fisher Fishkin and Zheng Zhang
Country: France

Abstract: ResOurceS for Endangered languages Through TranslAted texts (ROSETTA) is a collaborative and interdisciplinary project. Very much as the Rosetta stone helped decipher the demotic and hiero- glyphic scripts thanks to the presence of the Greek translation, the Rosetta project intends to preserve contemporary endangered languages and assist with their survival through translation. Our project puts to use the extant translated versions of a single literary text into a number of endangered languages over a rather long period of time. A first experiment was conducted on existing Basque translations of the well-traveled American novel “Adventures of Huckleberry Finn”.

Native language: روزيتا هو مشروع تعاوني ومتعدد التخصصات. بقدر ما ساعد حجر رشيد في فك رموز النصوص الديموغرافية والهيروغليفية بفضل وجود الترجمة اليونانية، يعتزم مشروع روزيتا على الحفاظ على اللغات المهددة بالإنقراض والمساعدة في بقائها من خلال الترجمة. يهدف مشروعنا إلى استخدام الترجمات الموجودة لنص أدبي واحد في عدد من اللغات المهددة بالإنقراض على مدى فترة زمنية طويلة. أجريت أول تجربة في هذا المشروع على ترجمات الرواية الأمريكية مغامرات هاكلبري فين المتوفرة في اللغة الباسكية المهددة بالانقراض.

P.2.20: Towards ASR that recognises everyone in a country with no spoken standard

Authors: Benedicte Haraldstad Frostad

Country: Norway

Abstract: Norwegian has many dialects, two written and no spoken standard. Speakers are used to linguistic diversity, and dialects are strong identity markers. Changing one's dialect is associated with identity loss. The extra costs, need for specialised expertise in spoken Norwegian and lack of suitable lexica and speech data sets complicate the development of ASR-products for this language community. This poses a democratic problem as public institutions automatise dictation and integrate ASR as a means for interaction. The Language Council is initiating innovative projects to improve ASR for Norwegian and minority languages in Norway and wishes to exchange ideas and experiences.

Native language: Norsk har to offisielle skriftnormer, mange dialekter og ingen offisiell uttalenorm. Språkbrukerne er vant til språklig mangfold, og dialekter er sterke identitetsmarkører. Å endre på dialekten sin forbindes med identitetstap. De ekstra kostnadene, den spesialiserte ekspertisen i norsk talespråk som trengs, og mangelen på passende leksika og taledatasett vanskeliggjør utvikling av talegjenkjenningsprodukter for dette språksamfunnet. Det er et demokratisk problem ettersom offentlige institusjoner innfører automatiske dikteringsverktøy og integrerer talegjenkjenning i sine kommunikasjonskanaler. Språkrådet har tatt initiativet til nyskapende prosjekter skreddersydd for å øke kvaliteten på talegjenkjenning for norsk og minoritetsspråk i Norge og ønsker å utveksle idéer og erfaringer.

P.2.21: Komi Latin-Alphabet Letters Not Found in Unicode

Authors: Jack Rueter

Country: Finland

Abstract: The two literary languages Komi-Permyak and Komi-Zyrian have used numerous alphabets and orthographies from the 13th century on. There are approximately six years of extensive publications in without recognizable texts. The typography is considered inconsistently poor, and therefore it can be categorized as a transitional alphabet. The term transitional alphabet, in turn, means that Unicode characters from mixed ranges can be used to satisfy many of the missing letters. The poster will illustrate missing letters in the Latin range with discussion and derive the minimal alphabetical requirements for the documentation of these two Uralic languages of Russia.

Native language: Кыкнан коми кывъяслӧн татчӧдз гижанногныс уна вӧлісны дас коймӧд нэмсянь. Оз тырмыны UNICODE-ын шыпасъяс квайт кымын волӧн гижӧдъяслы, кодъяс вывті коланаӧсь 1930-ӧд вояссянь. Шуӧны типографияыс пӧ омӧль, да та вӧсна колӧ чайтыны, мый вужӧдана тайӧ латиницаӧн гижанногыс. Вужӧдана – тайӧ позяс вӧдитчыны быдлаысь шыпасъястӧ – латиницаысь да кириллицаысь тшӧтш. Петкӧдлам став тырмытӧм шыпасъяс, кодъяс пиысь медся этша, мый колӧ, медым Россияса на кыкнан урал кывъяслы 6 во лыддянтортӧ сетны электроннӧй ногӧн.

P.2.22: Developing technologies for low-resource Uralic languages: Case studies on Saami and Komi varieties

Authors: Niko Partanen, Michael Rießler and Thierry Poibeau

Country: Finland

Abstract: The Uralic languages are spoken in northern Eurasia, and almost all of them are endangered. Language technology can play a major role in documenting and describing these languages better, and in making related workflows faster and more efficient. However, applying modern methods effectively in this context remains a challenge.

We have developed language technology for Komi and Saami, with a focus on a low-resource scenario. Besides providing an overview of this work, we detail what we see as the main challenges. Although we focus on individual languages, our experiences also translate to the wider situation of endangered languages outside Eurasia.

Native language: Uralilaisia kieliä puhutaan laajalla alueella Pohjois-Euraasiassa, ja valtaosa niistä on uhanalaisia. Kieliteknologialla voi olla merkittävä rooli näiden kielten kuvaamisessa ja dokumentaatiossa, erityisesti tehden näihin toimiin liittyvistä käytännöistä tehokkaampia ja nopeampia. Kieliteknologian nykysovellusten hyödyntämisessä tässä kontekstissa on silti runsaita ratkaisemattomia haasteita.

Työryhmämme on kehittänyt kieliteknologiaa saamelaiskielille ja komille, erityisesti tilanteeseen, jossa käytettäviä resursseja on vähän. Kuvaamme aiemmin tehdyn työn sekä keskeisimmät ongelmakohdat. Vaikka keskitymme yksittäisiin kieliin, ovat kokemuksemme sovellettavissa myös muihin uhanalaisiin vähemmistökieliin oman alueemme ulkopuolella.

P.2.23: Understanding culture and society with the language resources and tools offered through the CLARIN Research Infrastructure

Authors: Maria Eskevich and Franciska de Jong

Country: Netherlands

Abstract: Europe's Common Language Resources and Technology Infrastructure (CLARIN) aims at making language resources and tools from all over Europe and beyond accessible for research purposes through a single sign-on platform. CLARIN supports academic researchers, students, journalists and citizen-scientists interested in digital language resources, such as parliamentary records, social media data, newspaper archives and spoken corpora, and also functions as a knowledge sharing ecosystem. CLARIN adheres to the FAIR data principle. Open access to digital language resources that capture social and cultural diversity can help advance the social sciences and humanities at large.

Native language: De Europese onderzoeksinfrastructuur CLARIN (Common Language Resources and Technology Infrastructure) maakt taaldata vanuit de hele wereld en digitale analysetools voor taal via een 'single sign-on' platform toegankelijk voor onderzoeksdoeleinden. CLARIN ondersteunt academische onderzoekers, studenten, journalisten en citizen-scientists die gebruik maken van taalmaterialen (zoals parlamentaire verslagen, social media data, krantenarchieven en gesproken corpora), en functioneert tevens als een ecosysteem voor het delen van kennis. CLARIN is gebaseerd op de principes van FAIR data. Digitale taalmaterialen die vindbaar en toegankelijk zijn en hun inherente sociale en culturele diversiteit zijn van belang voor het domein van de sociale en geesteswetenschappen in brede zin.

P.2.24: A Multimodal Database of Russian Sign Language
Authors: Alexey Karpov, Ildar Kagirov, Dmitry Ryumin and Alexander Axyonov
Country: Russian Federation

Abstract: We present a multimodal database of Russian sign language (RSL) - TheRuSLan. It includes lexemes from RSL within one subject area demonstrated by 14 informants that were recorded with Kinect 2.0 sensor in FullHD video, infrared and depth map modes. RSL has an official status in the Russian Federation, and over 120K deaf people in Russia and some neighbour countries use it as their main language of spoken communication. RSL has no written system, poorly described and has very few electronic resources. We compare our database with other RSL corpora, and formulate some basic principles of gesture lexeme description.

Native language: Мы представляем многомодальную базу данных (тезаурус) русского жестового языка (РЖЯ) - TheRuSLan. База данных включает демонстрации лексем РЖЯ, относящиеся к одной предметной области и показанные 14 информантами. Данные были записаны при помощи устройства Kinect 2.0 в формате FullHD, в инфракрасном диапазоне и в режиме карты глубины. РЖЯ является одним из официальных языков общения на территории Российской Федерации, его носителями являются свыше 120 тыс. людей в России и сопредельных странах. РЖЯ не обладает системой письменности, недостаточно описан и имеет очень мало электронных ресурсов. Мы сравниваем нашу базу данных с другими корпусами РЖЯ и формулируем основные принципы описания жестовых лексем.

P.2.25: Sámi languages
Authors: Mikkel Rasmus Logje
Country: Norway

Abstract: Sámi languages are defined as a branch of the Uralic language family, and are traditionally spoken in an area stretching from central Norway and Sweden, through northern Norway, Sweden and Finland, to the Kola Peninsula in Russia. Today there are altogether 9 Sámi languages which are more or less mutually unintelligible, especially those that are geographically distant. The traditional boundaries between Sámi languages do not follow the national boundaries. The number of language users varies from one language to another. The largest language is Northern Sámi (est. 20.000–40.000 users). All Sámi languages are minority languages in the respective countries.

Native language: Sámegielat gullet urálalaš gielaide, ja daid árbevirolaš hupmanguovlu gokčá guovlluid Gaska-Ruota ja Gaska-Norgga rájes, Davvi-Norgga, Davvi-Ruota ja Davvi-Suoma bokte, gitta Guoládatnjárgii Ruoššas. Dál gávdnojit oktiibuot 9 sámegiela, ja gielaid gaskka leat unnit eanet erohusat, eandalii daid gielaid gaskka mat leat guhkkálaga. Sámegielaid árbevirolaš hupmanguovllut eai čuovo riikkarájiid. Leat erohusat daid iešguđetge gielaid geavaheddjiid logus. Stuorámus giella lea davvisámegiella (sullii 20.000–40.000 geavaheaddji). Buot sámegielat leat minoritehtagielat dáin guoskevaš riikkain.

P.2.26: LT Data Free for All
Authors: Marko Tadić and Tamás Váradi
Country: Hungary

Abstract: Language technology crucially depends on large amounts of texts. Digitally published text is a natural source for fast production of the fundamental language resources – corpora. However, clean, openly and freely available texts are difficult to come by. Even national languages suffer from scarcity of quality language data. We are presenting a project that can serve as a role model for the collection of large monolingual corpora for under-resourced languages. The approach could be applicable to any linguistic community that publishes legislative texts in their own language in digital form, to quickly build very big corpora

Native language: A nyelvtechnológia számára alapvető fontosságú az óriás mennaiségű szövegek elérhetősége. A digitálisan publikált szövegek kézenfekvő forrásai az alapvető nyelvi erőforrások, a korpuszok gyors előállításának. Azonban a tiszta, nyílt és ingyenesen hozzáférhető szövegeket nehéz beszerezni. Még hivatalos nemzeti nyelvek is szenvednek a jó minőségű szövegek hiányától. Bemutatunk egy olyan projektet, amely mintául szolgálhat arra, hogy miképpen lehet nagyméretű egynyelvű korpuszokat építeni erőforráshiányos nyelveken. A módszer minden olyan nyelvi közösség esetében használható, amely digitális alakban teszi közzé a saját nyelvén a jogszabályokat, melyekből hatékonyan lehet nagyon nagyméretű korpuszokat építeni.

P.2.27: Can we use a spoken Dialogue System to document Endangered Languages?
Authors: Jacqueline Brixey, Seyed Hossein Alavi and David Traum
Country: United States

Abstract: We investigate using a dialogue system to preserve endangered languages, and the viability of a multilingual dialogue system to generate a general use corpus of audio responses in . We introduce DAPEL (Dialogue APP for Endangered Languages). DAPEL elicits responses from speakers of endangered languages by having a conversation with them. We conducted a pilot user study to examine the efficacy of using an automated system like DAPEL versus a human interviewer. We also studied the effects of engaging in small-talk in a different language in between recording prompts for the target language.

P.2.28: Technologies for Endangered Languages: The Case of the Languages of Sardinia

Authors: Adrià Martín-Mor

Country: Andorra

Abstract: This poster shows the impact that technology may have on endangered languages, with a focus on Sardinian, one of the five native languages of Sardinia, according to the regional law. Specifically, it presents an example of how technology can be used to translate online texts, localise digital products and develop language resources. By resorting to free-licensed products, the output of these efforts can be re-used to generate further resources that, in turn, help increase the amount of texts generated.

Native language: Custu poster ammustrat s'impatu chi sa tecnologia podet tènnere in is limbas in perìgulu, e mescamente in sa limba sarda, una de is chimbe limbas nativas de Sardigna segundu s'istatutu regionale. In s'ispetzìficu, si presentat un'esempru de comente sa tecnologia podet èssere impreada pro bortare testos in lìnia, localizare produtos digitales e isvilupare resursas linguìsticas. Tràmite su sèberu de produtos cun lissèntzias lìberas, su resurtadu de custu traballu podet èssere aprofitadu pro generare àteras resursas chi, a su turnu suo, podent agiudare a crèschere sa cantidade de testos generados.

P.3.1: Challenges for language technologies in Ayapaneco
Authors: Jhonnatan Rangel
Country: France

Abstract: There are currently 577 critically endangered languages in the world, making up almost 10% of all languages. These languages are only spoken by a few elder speakers and are technologically low-resourced. Numde ‘oode or Ayapaneco is one of these languages, spoken by less than 11 elders in southern Mexico. Ayapaneco, like other critically endangered languages, poses various fundamental challenges including the annotation bottleneck that limits the scope of its documentation, preservation, reclamation, revitalization and utilization in language technologies. This poster addresses the challenges Ayapaneco confronts as it is vanishing before our eyes.

Native language: En el mundo hay 577 lenguas en muy alto riesgo de desaparición, constituyendo casi el 10% del total. Estas las hablan algunos adultos mayores además de que tienen pocos recursos tecnológicos. Numde ‘oode o ayapaneco, hablada por menos de 11 adultos mayores en el sureste mexicano, es una de estas. El ayapaneco, como otras lenguas en muy alto riesgo de desaparición, plantea retos fundamentales como el cuello de botella de anotación que limita las posibilidades de su documentación, mantenimiento, recuperación y uso en tecnologías del lenguaje. Este poster aborda los retos que enfrenta el ayapaneco al borde de la desaparición.

P.3.2: Mainumby: computer-assisted Spanish-to-Guarani translation
Authors: Michael Gasser
Country: United States

Abstract: Technology plays an important role in the daily work of the modern translator. However, computer-assisted translation (CAT), like machine translation, relies on extensive bilingual corpora, which are only available for a small minority of the world's languages. This poster presents a framework for the development of rudimentary CAT systems for translation into languages with limited resources and the compilation of bilingual corpora as a side-effect of the systems' use by translators. The framework has been implemented in Mainumby, a web application for CAT from Spanish to Guarani, the majority language of Paraguay.

Native language: Tuichaite mba’e niko ñe’ẽasahára ko’ağagua rembiapo pa’ũme mohendaha ha opaichagua apopyre jeporu. Upéicharamo jepe umi pojoaju, oñemboheráva ñe’ẽasa mohendaha ñepytyvõ rupive (ÑMÑ), oñemopyrenda ñe'ẽkõi retépe, ha tete kakuaaitéva ojejapo mbovymi ñe'ẽmente. Ko jehaipyre ohechauka mba’éichapa ojejapokuaa ÑMÑ ñe’ẽnguéra oñemomichĩvape ğuarã, ha omombe'u Mainumby, peteĩ apopyrã ÑMÑ oñemongakuaáva hína ojejapo porãve hağua ñe’ẽasa kastellánogui guaraníme. Upe pojoaju ikatu oipytyvõ porã tapicha ñe’ẽasahárape ha avei ombyaty umi ñe’ẽ ñe'ẽkõi rete.

P.3.3: Baby Quechua robot
Authors: Maximiliano Duran
Country: France

Abstract: A robot using artificial intelligence and a comprehensive set of linguistic resources may help to preserve Quechua. It may help in M.T of scientific, and cultural French documentation into Quechua. Written documentation, is essential to keep this language alive. I have been working on such a robot, for several years. I named it Yachaj/expert. I will show the first stage of this project: Baby Quechua Robot. Its functions are Automatic conjugation, lexical queries of Quechua-FR-SP; elementary spelling checking; and transliteration (alpha version) of texts written in the official spelling of Cuzco, Ecuador or Bolivia to that of Ayacucho and vice-versa.

Native language: Sinchillatañam Peru kitipi runasimi chintiramun kay ñawpaq pachak watallapi. Chaymi nichiwanchik: cheqappunim runasimiqa wañunayaypaq kachkan! Utjayllañam ima kutirichiykunatapas rurananchik ama kay simi wañunampaj. Allin qispichisqa, allin yachachisqa, llapan rikchaq runasimi cheqap-yachaykunawan kikin-ruraqqa yanapakuwanchikmanpunim runasimi unanchaypi. Chaymi ñuqa, kay ñawpaj qanchis watakunapi “yachachichkani” runasimita huk kikin-ruraqta. Paymi yanapakunqa mana-sasa runasimi yachachiykunata. Paytaqmi, allintaña puqurquspaqa. Payqa yanapawananchik punim Fransepi, Castellanopi qellqakunata runasimiman tikraipi, chayna achkallaña runasimipi qellqasqa taqekuna kanampaq. Cheqap-yachaymanta qellqakuna, willakuy-yachaymanta, llimpi-taki-yachaykuna qellqakuna achkallaña runasimipi qellqasqa taqekuna rikurinampaq. Chaynam michasun mana runasimi wañunapaq.

P.3.4: On the development of the Mexican Languages Parallel Corpus
Authors: Cynthia Montaño, Gerardo Sierra Martínez and Gemma Bel-Enguix
Country: Mexico

Abstract: The project we present is called Mexican Languages Parallel Corpus (CPLM) and its main goal is to contribute to development of NLP for low-resources Mexican languages. The CPLM consist of two modules: core module and subcorpus of religious and political texts module. The core module currently comprises 6 linguistics groups from 3 linguistics families; Mayan: Yucatec Maya and Ch’ol; Otomanguean: Mazatec, Zapotec and Otomí; Uto-Aztec: Nahuatl. The STRyP comprises 83 translations of the New Testament and 11 translations of three types of texts. The STRyP comprises a wide range of languages.

Native language: El proyecto que presentamos se llama Corpus Paralelo de Lenguas Mexicanas y su objetivo principal es contribuir al desarrollo de PLN para las lenguas de bajos recursos. El CPLM se compone de dos módulos: el módulo nuclear y el módulo de subcorpus de textos religiosos y políticos (STRyP). El módulo nuclear contiene actualmente seis grupos de tres familias lingüísticas; maya: maya yucateco y ch’ol; otomangue: mazateco, zapoteco y otomí, y yutoazteca: náhuatl. El STRyP se basa en 83 traducciones del nuevo testamento y once traducciones de tres tipos de textos. El STRyP se compone de un amplío rango de lenguas.

P.3.5: Project: Endless Oaxaca Multilingual
Authors: Tajëëw Díaz
Country: Mexico

Abstract: The Endless Oaxaca Multilingual project is an interdisciplinary project to bring computer equipment with the Endless operating system to indigenous communities in Oaxaca Mexico that have diverse content in the indigenous languages spoken in each community. The contents are mainly books for first readers, with the perspective of developing desktop applications in the coming months that help teachers to teach the indigenous language of the community. Rural communities in Oaxaca have very limited internet connectivity, so we plan to focus on content that may be available off line.

Native language: El proyecto de Endless Oaxaca Multilingüe es un proyecto interdisciplinario para llevar equipos de cómputo con el sistema operativo Endless a comunidades indígenas de Oaxaca México que tengan diversos contenidos en las lenguas indígenas que se hablan en cada comunidad. Los contenidos son principalmente libros para primeros lectores, con la perspectiva de desarrollar en los próximos meses aplicaciones de escritorio que ayuden a los profesores a la enseñanza de la lengua indígena de la comunidad. Las comunidades rurales en Oaxaca tienen conectividad a internet muy limitad, por lo que planeamos enfocarnos a contenido que pueda estar disponible sin necesidad de internet.

P.3.6: Large-scale audio-recordings to study infant language acquisition
Authors: Camila Scaff, Marvin Lavechin and Alejandrina Cristia
Country: France

Abstract: Studies of individual and socioeconomic variation in North America suggest that infant-directed speech quantities determine children's language advancement, inspiring interventions to get parents to talk more to their child. In this context, day-long audio-recordings analysed with proprietary software trained on American data are increasingly used to measure children's input and production, but there is little research on how fair this technique is to other languages and cultures. We present results from 10 Tsimane' children and their families (>270h audio, ~5h hand-annotated). Identification Error Rates averaged 62% (range 0-100%), inviting further work on open source diarization solutions that are retrainable.

Native language: Numerosos estudios sobre la variación individual y socioeconómica en América del Norte sugieren que las cantidades de habla dirigida a los bebés determinan el avance del lenguaje de los niños, lo cual ha inspirado intervenciones para que los padres hablen más con sus hijos. En este contexto, las grabaciones de audio de día completo analizadas con un software patentado entrenado en datos estadounidenses se utilizan cada vez más para medir la producción de los niños y cuanto se les habla, pero hay poca investigación sobre cuán justa es esta técnica para otros idiomas y culturas. Presentamos resultados de 10 niños Tsimane' y sus familias (> 270h de audio, ~ 5h anotadas a mano). Las tasas de error de identificación promediaron el 62% (rango 0-100%), invitando soluciones de diarización con código abierto y re-entrenable.

P.3.7: Nierika Red Social para aprender y enseñar una lengua indígena
Authors: Vania Ramírez
Country: Mexico

Abstract: NIERIKA is a niche social network on development, which is founded on the objective to collaborate with and support the preservation of all the Mexican indigenous languages. This unique platform enables users to present and create their publications which contains original indigenous linguistic data and it's closest translation into modern-day Spanish that does justice to the original meaning, this in turn promotes the idea of gathering linguistic data for investigative and research purposes. Nierika intends to create a digital community for members who want to share, learn, study and preserve the long lost linguistic treasures of Mexico.

Native language: NIERIKA es una red social en desarrollo, que se funda en el objetivo de colaborar y apoyar la preservación de todas las lenguas indigenas mexicanas. Esta plataforma única permite a los usuarios presentar y crear publicaciones que contienen datos originales de las lenguas indígenas y su traducción más cercana al español moderno tratando de ser fiel al significado original, esto a su vez permite recopilar datos lingüísticos con fines de investigación y estudio. Nierika tiene la intención de crear una comunidad digital para los miembros que desean compartir, aprender, estudiar y preservar los tesoros lingüísticos perdidos de México.

P.3.8: PRESERVING INDIGENOUS LANGUAGES IN SOUTH AND CENTRAL AMERICA BY LEVERAGING OPEN LICENSING AND TECHNOLOGY
Authors: Purvi Shah
Country: India

Abstract: StoryWeaver is a digital platform with 17,000+ free storybooks in 200+ languages, including 18 South and Central American languages, of which 11 are Indigenous. This poster presentation highlights how StoryWeaver's initiatives and technology catalyse the revitalisation of Indigenous languages in this region: Archiving endangered languages through storybooks - the first books published in Chocholteco in over a decade Building a sustainable self-publishing model, empowering local communities to print books in places like Oaxaca, where content creation is highly regulated Supporting communities of practice and building content repositories in 'bridge' languages like Spanish which aid translations in Indigenous language

Native language: 'StoryWeaver' es una plataforma digital con 16,000 libros gratuitos en 200 idiomas. Incluyen 18 idiomas sur y centroamericanos, de los cuales 11 son indígenas. Esta presentación destaca cómo las iniciativas y la tecnología de StoryWeaver aceleran la revitalización de las lenguas indígenas:

1. Archivar lenguas en peligro de extinción a través de libros y salvarlas de la extinción
2. Construir un modelo de autopublicación que empodera las comunidades locales para imprimir libros en lugares donde la creación de contenido está altamente regulada
3. Apoyar a las comunidades a construir repositorios en idiomas como Español, que ayudan a las traducciones en lenguas indígenas

P.3.9: Comunidad Elotl. Language Technologies for Mexico's Indigenous Languages
Authors: Ximena Gutierrez-Vasques and Victor Mijangos
Country: Mexico

Abstract: Comunidad Elotl gathers a group of enthusiasts that share the interest of generating language technologies for the languages spoken in Mexico. So far, our projects have focused in the recollection of parallel corpora, building accesible web search interfaces for these corpora and in the research and development of Natural Language Processing (NLP) Techniques for building taggers.

In this poster we summarize our main contributions, moreover, we highlight some of the main challenges that arise when dealing with these low-resource languages from a computational perspective.

Native language: (Nahuatl) Nechikol Elotl kichiua tein mo tekipachoa maj onka tajtol in amantekayotl tech Mexiko tajtolmej. Axkan sayo tik sentilia “corpus paralelos”, uan tlatemoa ika in tekit, no tik amatemoa uan tikchiua “Procesamiento del Lenguaje Natural (NLP)“. Itech in amatl tik nextia to tekit, no mo nextia ken oui etoke in tajtolmej itech in tonalmej.

P.3.10: Language and Landscape: Hiking and Documenting the Chatino Language of San Juan Quiahije
Authors: Emiliana Cruz
Country: Mexico

Abstract: A few remaining elder speakers of the San Juan Quiahije Chatino language (Oaxaca, Mexico) have unique command of specialized words, expressions, and grammatical features that relate to local landscapes and nature specific activities. In a moment marked by rapid decline of indigenous languages, this area of language undergoes swifter deterioration than any other. This poster displays a methodology that aids in the study of landscape specific language. Documentation and dissemination of collective knowledge alongside use of specialized place-based expressions highlights scholarly-community collaboration in indigenous language revitalization, and outlines my journey on foot with inhabitants of the San Juan Quiahije landscape.

Native language: waC tiC chiqH qaJ ntenB tqaG xiA tyinJ noA tiC jlyoH riqC neG saA skaI naF sqwiJ neqC xqoF. LaC qaE xqanE naF, qoE neC waC ndyiH snaC ndyiE chaqF jnyaJ. NaqG nyiA qyanH chaqF tiC chiqH qaJ ntenB noK tiC jlyoH riqC naF noJ yqwiJ tiC sqneE. LoA ktyiC reC ktsanH chinqH qwanK noK qneI waG jnyaF chaqF tiC xnyiJ tyqiC ntenB noK tiC jlyoH riqC saA skaI chaqF. tyonC noE ngaJ waG noA qneI waG jnyaF qinF ranF, qneJ waG jnyaF chaqF jaA tyiI chaqF jnyaJ.

P.3.11: Resources and digital materials in Mexico’s indigenous languages
Authors: Luis Flores Martínez
Country: Mexico

Abstract: In Mexico, 68 indigenous languages (LI) are spoken which are at risk of disappearance. The use of ICTs can be an ally for LIs by allowing them to increase their visibility, promoting their use, teaching, and learning. A multidisciplinary group of LI speakers created a platform on Facebook (@lenguasweb) intending to raise awareness of the linguistic richness, as well as teach illiterate speakers the written form of their own language. We hope to be a motive, especially for young people, and contribute to the preservation, dissemination, and revitalization of the minority languages of Mexico and the world.

Native language: Tsakam dhuchlab (Tének de San Luis Potosí, México) Ti Labtóm Tsabál ajyamej óx inik laju waxik i Tének kawintaláb, po axe'chik yab exladh, ani wa'ats i atiklabchik axi yabáts in le' kin eyendha'. Jaxtam jún kubél i atikláb axi i kawnál i Tének káwintal u junkun abal ki ts'ejka' jún i xeklek ti Facebook axi in jamat bij játs @lenguasweb, taná' i tejwa'medhál junchik i káw abal pilchik i atikláb kin exla'chik ti waw i kwentaj, po jayej abal i juntal Tének kin exla' jant'ini' tu dhucháb i káwintal; axé' jayej pel abal ki edhanchij in tsaláp pil i juntal ani ki ela' ti ébtsolom jun i jolataláb abal wawá' ani i káwintal.

P.3.12: Ayöök, México
Authors: Marco Martinez
Country: Mexico

Abstract: Kumoontun was developed out of necessity in order to preserve the Ayook language (a variant of Mixe from the Totontepec area) by taking advantage of digital media and platforms. It is a new and different experience because the Kumoontun App was created is shared directly with the communities by means of workshops for children, youth and the community at large. Today we share the experiences, challenges and achievements of this project only months after its creation.

Kumoontun App for IPhone https://cutt.ly/FeOlmPQ

Kumoontun App for Android https://cutt.ly/peOlTZb

Native language: Yi Kumoontun wä’ä yiwe tsyöönt’iik jats yiwe yi ayöök juu' yak’kojtsp Anyiköjmtsoj yajk’kojtswijï’t yak’kojtswa’atsï’t mëët yi tuköjtsïn, tonpäjtïn jats mëët jomajatsy yi ayöök wya’kxtïk. Pi’k’önïkta, waa’tyëjkta, kiixtë’ëxta jats kajpïn jayïta ananyijoma mëët ëëtse ntun. Xyam ëëtse’n’awanat wintsowe ja winma’ayïn myiijn jats yak’ukwaajny, tyëjxyiwe winköpk jats yiwe yi aa’ayöök ntöönkimtat.

Kumoontun App IPhone https://cutt.ly/FeOlmPQ

Kumoontun App Android https://cutt.ly/peOlTZb

P.4.1: African Wordnet – digital documentation and preservation of indigenous knowledge
Authors: Sonja Bosch and Marissa Griesel
Country: South Africa

Abstract: Indigenous knowledge concepts in isiZulu, collected from a variety of sources such as monolingual and bilingual dictionaries can be transformed from alphabetically ordered entries into a hierarchical wordnet structure as a set of relations. The ensuing synsets can further be enriched and lexical gaps filled with information from other cultural resources to transcend the physical limitations of such traditional sources by including definitions, usage examples, pictures and dialect information. The multilingual African Wordnet can be used as important tool in the globalisation of Africa's indigenous knowledge systems, thereby contributing to language empowerment through revitalization of culture and knowledge

Native language: I-African Wordnet - imibhalo eyidijithali kanye nokugcinwa kolwazi lwendabuko Kulokhu okwethulwayo kuzovezwa ukuthi amagama esiZulu awulwazi lwendabuko, aqoqwe eqhamuka emithonjeni enhlobonhlobo enjengezichazimazwi ezinolimi olulodwa nezinezilimi ezimbili, angaguqulwa asuke emagameni afakwe ahleleka ngokwe-alfabhethi, abe yisakhiwo samagama ahlelwe ngokwamazinga ehlukahlukene saba yiqoqo eliveza ubudlelwane obukhona phakathi kwawo. Amaqoqo amagama amqondofana (synsets) atholakala kulokhu, aphinde athuthukiswe kuvalwe namagebe aphathelene namagama asolimini ngemininingwane eqhamuka kwezinye izinsiza zezamasiko ukuze kugwenywe ukungapheleli kwemithombo leyo ejwayelekile ngokuthi kufakwe izincazelo, izibonelo zokusetshenziswa kwawo kanye nemininingwane yolimi olusetshenziswe. I-African Wordnet enezilimi eziningi ingasetshenziswa njengethuluzi elibalulekile ekusebenzeni kumazwe ngamazwe kwezinhlelo zase-Afrika zolwazi lwendabuko, ngalokho iphonse itshe esivivaneni ekunikezweni kwezilimi amandla ngokuvuselelwa kwamasiko nolwazi.

P.4.2: Automated Speech Segmentation: Example of an African Language
Authors: Brigitte BIGI
Country: France

Abstract: Speech segmentation is the process of identifying boundaries between speech units in the speech signal and determining where in time they occur. Linguistic resources of the target language should be defined: a lexicon (the words to be recognized), a word dictionary (their pronunciations as a sequence of phonemes), an acoustic model (a stochastic representation of input waveform patterns per phoneme).

SPPAS software tool implements language-and-task-independent algorithms. This multilingual approach was applied to the african langage Naija (Nigerian pidgin). We developped language resources for a tokenizer, an automatic speech system for predicting the pronunciation of the words and their segmentation.

Native language: La segmentation de la parole consiste à identifier les unités dans le signal de parole et à déterminer où celles-ci se produisent dans le temps. Des ressources linguistiques de la langue cible doivent être définies : un lexique (les mots à reconnaître), un dictionnaire de mots (leurs prononciations en tant que séquence de phonèmes), un modèle acoustique (une représentation stochastique par phonème).

L'outil logiciel SPPAS implémente des algorithmes indépendants du langage et des tâches. Cette approche multilingue a été appliquée au langage africain Naija (pidgin nigérien). Nous avons développé des ressources linguistiques pour un tokenizer, un convertisseur graphème-phonèmes et leur alignement avec le signal.

P.4.3: Establishing Sustainable Infrastructures for African Languages
Authors: Z Steyn
Country: South Africa

Abstract: The South African Centre for Digital Language Resources (SADiLaR) is a new research infrastructure (RI) set up by the Department of Science and Innovation (DSI) forming part of the new South African Research Infrastructure Roadmap (SARIR)

The centre runs two main programmes. A digitisation programme and a Digital Humanities (DH) programme. The focus of the poster will show the different elements of establishing the digitisation programme, which focusses on the creation text, audio and multimodal datasets as well as the development of NLP tools and software for the 11 official languages of South Africa.

Native language: Die Suid Afrikaanse Sentrum vir Digitale Taal Hulpbronne (SADiLaR) is 'n nuwe navorsingsinfrastruktuur befonds deur die Departement van Wetenskap en Innovasie en maak deel uit van die Suid Afrikaanse Navorsingsingsinfrastruktuur Padkaart (SARIR)

Die sentrum huisves twee programme. 'n Digitalisering en 'n Digitale Humaniora-program. Die plakkaataanbieding sal fokus op die digitaliseringsprogram, spesifiek op die samestelling daarvan wat behels die skep van teks, klank en ander multimodale datastelle sowel as die ontwikkeling van NLP programme vir die 11 amptelike tale van Suid Africa.

P.4.4: A South African Corpus of Multilingual Code-switched Soap Opera Speech
Authors: Febe De Wet, Ewald Van der westhuizen and Thomas Niesler
Country: South Africa

Abstract: We introduce a speech corpus containing multilingual code-switching compiled from South African soap operas. The corpus contains monolingual as well as code-switched examples of English, isiZulu, isiXhosa, Setswana and Sesotho speech. The last four are indigenous languages, all belonging to the Southern Bantu family. IsiZulu and isiXhosa are Nguni languages that, while distinct, are to some degree mutually intelligible and linguistically similar. The same applies to Setswana and Sesotho, which are Sotho-Tswana languages. The data contains both inter-sentential and intra-sentential code-switching. Intra-sentential code-switching occurs as alternation, insertion as well as intra-word switches.

Native language: Sethula i-corpus yenkulumo equkethe ukushintshwa kwekhodi yezilimi eziningi ehlanganiswe kuma-soap opera waseNingizimu Afrika. I-corpus iqukethe izibonelo zesiNgisi nesiZulu nesiXhosa nesiTswana nesiSuthu ezinolimi olulodwa kanye nezibonelo ezishintshile ikhodi. Ezine zokugcina ziyizilimi zomdabu, zonke zingabomndeni waseSouthern Bantu. IsiZulu nesiXhosa yizilimi zesiNguni, nakuba zihlukile, ngezinga elithile ziyaqondana futhi zifana ngohlelo. Kwenzeka okufanayo nesiTswana futhi nesiSuthu, eziyizilimi zesiSuthu-Tswana. Idatha iqukethe ukushintshwa kwekhodi okungaphandle kwemisho futhi okungaphakathi kwemisho. Ukushintshwa kwekhodi okungaphakathi kwemisho kwenzeka njengokushintshana (alternation), ukufakwa (insertion) kanye nokushintshwa ngaphakathi kwamagama (intra-word switches).

P.4.5: Corpora Mandeica: text corpora for Mande languages (West Africa)
Authors: Valentin Vydrin
Country: France

Abstract: "Corpora Mandeica" is a set of corpora of annotated written texts in languages of the Mande family, openly accessible in the Internet. All the texts in the corpora are provided with POS tags and French (eventually also English and Russian) glosses. The corpora are partly disambiguated; parallel subcorpora are also being developed. So far, there are corpora for four languages available on line: Bambara (more than 11 million words), Guinean Maninka (about 3,5 million words), Eastern Dan (about 460,000 words), Mwan (47,000 words). The corpora are accompanied by electronic dictionaries and electronic libraries. Further language corpora are planned.

Native language: Проект Corpora Mandeica представляет собой совокупность аннотированных корпусов письменных текстов на языках манде, находящихся в открытом доступе в Интернете. Все тексты аннотированы (снабжены частеречными пометами и французскими глоссами; отчасти также английскими и русскими). Для части текстов проведено снятие омонимии. Создаются также параллельные корпуса. К настоящему моменту доступны для поиска корпуса 4 языков: бамана (11 млн. слов), гвинейский маninка (около 3,5 млн.), восточный дан (около 460 тыс. слов), муан (47 тыс. слов). На корпусных сайтах вывешены электронные словари; имеются также электронные библиотеке для 3 языков. Планируется создание корпусов и для других языков семьи.

P.4.6: Missing link: A centralised digital archive for endangered languages of southern Africa
Authors: Kerry Jones
Country: South Africa

Abstract: Language endangerment and language loss is a worldwide phenomenon. As a result, the scramble to identify, document and preserve indigenous languages using digital technology has gained traction. The challenge we face in southern Africa, is the lack of a centralised digital archive for endangered languages. Currently, efforts are dispersed on various platforms, hosted by universities, non-government organisations or private collections, if digitised at all. In order to provide a holistic description of endangered and extinct languages in southern Africa, an online digital archive could centralise existing efforts, while creating opportunities for the digitisation of historical records and new digitised entries.

Native language: Gowaga ǁō‡oas !aorosasib tsî gowaga laris tsîra ge !hūbaib ǂhabase, harase a ǂansa ǃnae!khaira. ǁNā-amaga di ge loro ǁan‡gāsaben gowaga nēsi hâ texnoloxib lkha ôa‡ui, xoamâi tsi lkhaubas di lgūbade nēsi lgaisase ra ǂoaxa. Afrikab !khawagas ǃnâ da ra hō!â nausa ǁgoa‡uis ge lgui!nâxa digitel‡khanisâuǁgâugu !nuwusiba, ǁō‡oas !aorosasib ǃnâ mâ gowagu !aroma. Nē ǁaeb ai di ge lguilguibe dītsârode !kharaga!nâgu !harodi ai ra hōhe, universiteitdi tawa i ka hâ tama kara io, o ǂhanub !auga hâ ǂnûi‡gādi tawa, tamas ka io, lguilguibe khoen tawa - ǁnās ge hâna i ka digitelǁgaub ǃnâ a hōhe ǁkhā osa. Hoa !hariga !khō‡gā hâ lgaub ǃnâ da ka ǁō‡oas !aorosasib ǃnâ mâ gowagu tsî lnai ge ǁō‡oa gowagu Afrikab !khawagas digu tsîna a xoamâi ǂgao, o i ge kaise nî !gâi online digitel‡khanisâuǁgâuba kurusa. Nēti ī digitel‡khanisâuǁgâub ge lnai hâ sîsengu hoaga lhaolhao tsî lgui !khais tawa lgui nî !khōlgara, tsî nēs ǃnâ-u lgūlgarus !nae!khaidi xoalǁguigu tsî ka hâ lasa xoadi tsîna digitallgaub ǃnâ sâus di sîsen-i di daode nî ǁkhowa-am.

P.4.7: Using Citizen Linguistics to Empower Indigenous Communities
Authors: Christopher Cieri and Mark Liberman
Country: United States

Abstract: While Language Technologies promote digital linguistic diversity and community development through access to knowledge, such technologies rely upon datasets absent in most indigenous languages. The LanguageARC Citizen Linguist portal augments traditional sources of language data by empowering indigenous communities to create their own. Via brief, engaging tasks such as picture and video description, vocabulary elicitation and usage surveys that can be completed on a computer or smart phone, indigenous communities collect spoken or written data as appropriate, and augment it by transcribing, translating, judging grammaticality and annotating for use in technology development and language development.

P.4.8: Heuristic guided probabilistic graphic language modelling for morphological segmentation of isiXhosa

Authors: Lulamile Mzamo, Albert Helberg and Sonja Bosch

Country: South Africa

Abstract: The IsiXhosa Heuristics Maximum Likelihood Segmenter (XHMLS), an unsupervised isiXhosa segmenter, is evaluated. The study contributes use of isiXhosa word morphology heuristics as a guide to probabilistic graphical modelling (PGM) the segmentation of isiXhosa. Four guided PGMs with options for modified Kneser-Ney (mKN) smoothing are presented. XHMLS's boundary identification accuracy of 78.7% outperforms the benchmark Morfessor-Baseline's 77.2%, and shows an even better f1-Score, 68.0%, compared to Morfessor-Baseline's 48.9%, when modelled with circumfixing and smoothing. The study shows that better word segmentation performance could be achieved in the unsupervised morphological segmentation of isiXhosa if a representative and smoothed PGM is used.

Native language: I-IsiXhosa Heuristics Maximum Likelihood Segmenter (XHMLS), isicaluli-mbhalo sesiXhosa esingagadwanga, siyavavanywa. Igalelo loluphando kukusetyenziswa kwendlela amagama esiXhosa aguquka ngayo njengesikhokelo somFanekiso-mBoniso-Thuba (FBT) ekucaluleni isiXhosa. Ii-FBT ezikhokelweyo ezine, ezinokukhetha ukusebenzisa ugudiso lwe-Kneser-Ney elungisiweyo (mKN), ziyaboniswa. Inkcaneko yokukhomba imida yezimilo ye-XHMLS eyi-78.8% igqitha eyomgangatho-jikelele oyiMorfessor-Baseline, we-77.2%, kwaye igqithe ngakumbi ngenqaku le-f1, ngo-68.0%, xa ithelekiswa neyeMorfessor-Baseline engu-48.9%, xa inkokhelo izizimi-macala yaye igudiswe nge-mKN. Olu phononongo lubonisa ukuba ucalulo-magama lwesiXhosa olungcono lungafumaneka xa kusetyenziwa isicaluli-magama esingagadwanga se-FBT esisifuziseleyo isiXhosa sibe sigudiswe nge-mKN.

P.4.9: Radio-browsing in support of relief and development work in rural Africa

Authors: Astik Biswas, Febe De Wet, Herman Kamper, Raghav Menon, Thomas Niesler, Armin Saeb, John Quinn, Ewald Van der westhuizen and Emre Yilmaz

Country: South Africa

Abstract: In countries with well-established internet infrastructure, social media has become an accepted platform for voicing opinions. However, in some parts of Africa internet infrastructure is poorly developed, precluding the use of social media to gauge sentiment. Instead, community radio phone-in talk shows are used to voice views and concerns. Our contribution will introduce a radio browsing system that is intended to support relief and developmental programmes by the United Nations (UN). Browsing systems were developed to monitor community radio broadcasts for keywords related to specific topics such as natural disasters, disease outbreaks, or other crises.

Native language: Wadamadda leh kaabayaasha internetka ee sida wanaagsan loo aasaasey, warbaahinta bulshada waxay noqotey meel lagu aqbalo fikradaha codadka la dhiibto. Si kastaba ha ahaatee, qeyb ka mida kaabayaasha internetka ee Afrika ayaa si liidata loo horumariyey, iyadoo la sii saadaalinaayo adeegsiga warbaahinta bulshada si loo qiyaaso dareenka. Taa badal keed waa bandhigyada wada hadalka telifoonka telefishanka bulshada ayaa loo isticmaalaa in lagu dhawaaqo aragtida iyo shirarka. Waxqabadkeena ayaa soo bandhigi doona, nidaam raadiya raadyaha oo loogu tala galey in lagu taageero gargaarka iyo barnaamijyada horumarineed ee ay bixiso Qaramada Midoobey (UN). Nidaam baadhitaan ayaa loo sameeyey si loola socdo raadyaha idaacadda bulshadaee ereyada muhiimka ah ee la xidhiidha mowduucyada gaarka ah, sida masiibooyinka dabiiciga ah, cudurada faafa ama dhibaatooyin kale.

P.4.10: Analysis of Language Relatedness for the Development of Multilingual Automatic Speech Recognition for Ethiopian Languages

Authors: Martha Yifiru Tachbelie, Solomon Teferra Abate and Tanja Schultz

Country: Ethiopia

Abstract: In this poster, we present the analysis of GlobalPhone (GP) and speech corpora of Ethiopian languages (Amharic, Tigrigna, Oromo and Wolaytta). The aim is to select speech data from GP for the development of multilingual Automatic Speech Recognition (ASR) system for the Ethiopian languages. To this end, the phonetic overlaps among GP and Ethiopian languages have been analyzed. Moreover, morphological complexity of the GP and Ethiopian languages, reflected with high out of vocabulary rate and type to token ration, has been analyzed using training transcriptions. We also present baseline ASR performances for each of the GP and four Ethiopian languages.

Native language: በዚህ ፖስተር የምናቀርብላችሁ በግሎባል ፎን እና በአራት የኢትዮጵያ ቋንቋዎች (አማርኛ፣ ትግርኛ፣ ኦሮምኛ እና ወላይትኛ) የድምፅ ዳታ መካከል ያደረግነውን የማነጻጸር ጥናት ነው። የጥናቱ ዋና አላማ ከግሎባል ፎን የድምፅ ዳታ ውስጥ ለብዙ የኢትዮጵያ ቋንቋዎች ንግግርን ወደ ድምፅ የሚቀይር መተግበሪያ ለመስራት ጠቃሚ የሆነ የድምፅ ዳታ መምረጥ ነው። ስለዚህም በግሎባል ፎን እና በአራቱ የኢትዮጵያ ቋንቋዎች መካከል ያለውን የድምፅ መመሳሰል አጥንተናል። በተጨማሪም የቋንቋዎቹን ምእላዳዊ ውስብስብነት ለመረዳት እንዲቻል "Out of Vocabulary" እና "Type to Token Ratio" በማስላት ለማየት ሞክረናል። ለእያንዳንዱ የግሎባል ፎን እና የኢትዮጵያ ቋንቋዎች የተዘጋጁ ንግግርን ወደ ድምፅ የሚቀይሩ መተግበሪያዎችንና አፈፃፀማቸውን አሳይተናል።

P.4.11: Automatic Learning of a Phonological System: a Case Study on the Mboshi Language

Authors: Lucas Ondel and Lukas Burget

Country: Czech Republic

Abstract: Over the last decade, a lot of research focused on automatically learning basic acoustic units, a.k.a "pseudo-phones", for low-resource languages. In this work, we investigate the potential and the limits of this research on a real case scenario for documenting a low-resource language. We performed our experiments on Mboshi, an African language from the Bantu family. Results show that despite some progress, automatic learning from under-resourced languages remain a very challenging task and requires further research.

Native language: Au cours des dernières années, de nombreux travaux de recherche se sont concentrés sur l'apprentissage d'unités acoustiques, appelées "pseudo-phones", pour les langues peu dotées. Ce travail explore le potentiel et les limites de cette recherche dans un scénario réaliste de documentation d'une langue peu dotée. Nous avons mené nos expériences sur le Mbochi, une langue Africaine de la famille Bantoue. Les résultats montrent que, en dépit de progrès indéniables, l'apprentissage automatique à partir d'une langue peu dotée reste une tâche difficile et nécessite de plus ample recherches.

P.4.12: Current Status, Issues, and Future Directions for Ethiopian Natural Language Processing (NLP) Research
Authors: Seid Yimam and Chris Biemann
Country: Germany

Abstract: These days, the generation of resources (mainly text and speech) for many languages is dramatically increasing. However, high-resource languages such as English and low-resource languages such as Amharic, greatly differ on the amount of NLP components, tools and applications. In this poster, we will briefly discuss the state-of-the-art NLP research for Ethiopian languages. Then, the main bottlenecks that hinder the development of the required resources will be reviewed. Finally, we will point out best practices to solve current issues and indicate appropriate tools and models that can be easily adapted for low-resource NLP research, particularly for Ethiopian languages.

Native language: በአሁኑ ግዜ፣ ለብዙ ቋንቋዎች መረጃዎችን (በዋነኛነት የጽሑፍ እና የንግግር) ማግኘት እጅግ በጣም ቀላል እየሆነ እየሆነ መጥቷል። ሆኖም ግን እንደ እንግሊዝኛ ቡበዛት መተግበሪያ ያላቸውና እንደ አማርኛ ያሉ እጅግ በጣም ዝቅተኛ መተግበሪያ ያላቸው ቋንቋዎች፣ በተፈጥሯዊ የቋንቋ ቴከኖሎጂ (ተ.ቋ.ቴ - NLP)ግብዓቶች፤ መሳሪያዎች እና መተግበሪያዎች መጠን ሰፊ ልዩነት አላቸው። በዚህ ፖስተር ጽሁፍ፤ በመጀመሪያ የኢትዮያ ቋንቋዎች በተ.ቋ.ቴ ምርምር አሁን ያሉበት ደረጃ በጥልቀት ይብራራሉ። በመቀጠል ለኢዮጵያ ተ.ቋ.ቴ ምርምር እደገት እንቅፋት የሆኑ ዋና ዋና ክፍተቶች ይገመገማሉ። በመጨረሻም ፣ ወቅታዊ እና ተያያዥ ተግዳሮቶችን እንደት መፍታት እንደሚቻልእ፤ አሁን ላይ የሚገኙ የሌሎች የበለፀጉ የቋንቋ መተግበሪያዎችን እና ሞደሎችን የመተግበሪያ አጥረት ላለባቸው ቋንቋቆች(በተለይም ለኢትዮጵያን ቋንቋዎች) እንደት ማላመድና መጠቀም እንደሚቻል ይጠቆማል።

P.4.13: ACALAN: Platform for African Language Empowerment (PALE)
Authors: Martin Benjamin
Country: Switzerland

Abstract: The African Academy of Languages (ACALAN) is finalizing a proposal for a comprehensive platform for African languages in Cyberspace. The platform will serve four functions:

1. Information about ACALAN and African language policies and commissions
2. Information about African language research and characteristics
3. A hub for growth and dissemination of African linguistic data
4. A communications center for research and development on African languages

The goal is to produce an ever-growing central resource for scholars, policy-makers, students, and the public to learn about, contribute to, and benefit from knowledge regarding all African languages.

P.4.14: SCAnnAL – An Automatic Speech Corpus Annotator for African Speech Corpora
Authors: Moses Ekpenyong, Eno-Abasi Urua and Aniefon Akpan
Country: Nigeria

Abstract: Today, thousands of annotated speech corpora exist worldwide and demand for richly annotated corpora is fast growing, but the process accompanying the segmentation and labeling of corpora has slowed research progress for African languages due to the limitations of current annotation Toolkits to satisfy the challenges African speech systems present. We introduce SCAnnAL, a Toolkit for automatic speech annotation that automates the annotation process by accepting raw audio files, segments the waveforms and finally dumps labels into created segments. SCAnnAL is currently being refined for accuracy and is certain to put an end to the intractable procedure of speech annotation.

Native language: Mfịn ami, mme tọsịn adianañkpadia ikọ eba ke ekondo, ñko enekke eyem mme adianañkpadia ikọ ntom ke ise ikọ eti eti. Daña esaña esep enyVñ ewet anyịñ, anam nduuñọ ke mme usem Afrika anyọnyọọñ sia se ekama enam utom ado anana akeene ñkpọ ñnyan ubọk aabañake mme usem Afrika. Imiben SCAnnAL, anamidem akebe ñkpọ usep ñnyVñ ñwet anyịñ ikọ ndọñ ke ise ikọ iwọt. SCAnnAL akeme adidat utatañ ikọ, asep, anyVñ awet anyịñ ke mbaak ikọ. EsVk enanam utom ke SCAnnAL ma akan anọ nneke ibọọrọ anyVñ ayọ mfịna aasañake ke adisep ikọ ndon ke ikpeghe.

P.4.15: NTeALan - Artificial Intelligence, Development and Promotion of African National Languages
Authors: Elvis Mboning and Damien Nouvel
Country: France

Abstract: Among the emergent challenges that the African continent is currently facing is the problem of safeguarding and enhancing its cultural and linguistic heritage. Created in 2017, the NTeALan association has been working since then to implement intelligent technological tools for the digitization, promotion, development and teaching of African national languages. NTeALan wants to make these languages the pillars of social and technological development in Africa.

Native language: Í kété mìtíík mí mám má ńsòmblà íŋǔ hólós áfríkà í mà ŋgéŋ máná dì gwě, màhɔ̀l má má ntágɓéné í máhɔ́p més nì bìɓoŋɔ́l gwés, íŋǔ hàlà ɲén Ntealaɲ (ìɓòdòl 2017), tòhálá kíí à má sál ŋgándàk mú í ndzél ī, à ŋgí sálák nì láná léé: à ńtí bínóŋɔ́l bí mɔ́ndɔ́ bí bí ńhólá lɛ̀è dí nìígá, nì hólɔ̃s màhɔ́p més lòŋnì ndzél ì bìɓòŋɔ́l bí mɔ̀ndɔ́. hálà à gáhólá í nììgà nì hólós màhóp més . NTeALan à ńsòmból ɓɔ̀ŋ lɛ̀è màhɔ́p má áfríkà má ɓá ŋńɛ́m ú máhɔ̀l má bílɔ̀n gwés gwɔ́ bísóná.

P.4.16: Grappling with Opportunities and Challenges in developing Language Technologies for Under-resourced African Languages
Authors: Sunday Ojo
Country: South Africa

Abstract: Multilingualism and Diversity is the hallmark of African societal heritage, which can be exploited to foster unity in diversity, towards sustainable development. But challenges and opportunities abound in exploiting Language Technology (LT) potentials. Their under-resourcedness and rich contextual semantics features constitute a challenge, given that most of the LTs are validated using richly-resourced languages, such as English as test beds. There are also contextual issues bordering on paucity of internal capacity for exploiting the LT potentials. The poster draws from an on-going experience in working on LTs for some West and Southern African under-resourced languages, eliciting these challenges and opportunities.

Native language: Orisirisi ede ati Oniruuru jẹ aami pataki fun ohun-ini awujọpọ ti Ilu Afirika, eyiti a le lo lati ṣe isọdi iṣọkan ni oniruuru, si idagbasoke alagbero. Ṣugbọn awọn italaya ati awọn anfani pọ si ni ilokulo awọn agbara Imọ-ẹrọ ede. Wọn labẹ abinibi ati awọn ẹya Semantics ọrọ ọrọ ọlọrọ jẹ ipenija, fun pe opọlọpọ awọn Imọ-ẹrọ ede ni a fọwọsi pẹlu lilo awọn ede ti o lọpọlọpọ, bii Gẹẹsi bii awọn akete idanwo. Awọn ọran ti o tọ tun wa ni ilaja lori agbara ti agbara inu fun ilo awọn agbara Imọ-ẹrọ ede. Atẹjade naa fa lati iriri lilọ kiri ni sisẹ lori Imọ-ẹrọ ede fun diẹ ninu awọn ede ti Iwọ-oorun ati Gusu Afirika labẹ itusilẹ, ṣiṣe awọn italaya ati awọn aye wọnyi.

P.4.17: Towards the First Low-Resource French-Somali Machine Translation System
Authors: Houssein Ahmed Assowe and Fatiha Sadat
Country: Canada

Abstract: As part of an ongoing research on NLP for under-resourced languages, we have created several linguistic tools such as a French-Somali machine translation system dedicated to the journalistic domain, allowing to obtain quality translations from a bilingual corpus built by post-editing GOOGLE TRANSLATE results. The final users being the Somali and non-French speaking populations of the Horn of Africa. Furthermore, an aligned parallel corpus was constructed and evaluations were conducted by nine bilingual annotators who assigned a quality score to each segment of the corpus and corrected the post-editing results. We are following up with a neural French-Somali MT system.

Native language: Iyada oo qayb ka ah cilmibaadhista hadda jirta ee NLP-da oo loogu talagalay luqadaha aan aad uu kombuyutargaraynsanayn, waxaan abuurnay dhowr aalado luqadeed oo kala duwan sida nidaamka turjumaad mashiinka ee afka faransiiska iyo soomaliga kaas oo loo qoondeeyay aagga warbaahinta, isago oo oggolaanayno in la helo tarjumaad tayo leh oo laga sameeyay xogta laba-luqadlaha ah oo aan ka dhisnay dib-u-habeynta natiijooyinki turjumaadi uu horeysay ee GOOGLE TRANSLATE. Adeegsadayaasha aaladan ayaa ah dadka Soomaalida ee aan ku hadlin afka Faransiiska ee Geeska Afrika. Intaa waxaa sii dheer, xog laba-luqadle oo isbarbar socda ayaa la dhisay , iyo qiimeynna waxa ku sameeyay xogta sagaal tilmaameyaal laba luuqadoodla ah kuwaas oo siiyay dhibco tayo leh weer kasta oo ka mid ah xogka isla markaana saxay natiijooyinka dib-u-habeynta. Waxaan kale oo sameynay nidaamka turjumaad ee afka Faransiiska iyo Soomaaliga oo ku dhisan shabakadaha neerfaha ee aan dabiici ahayn .

P.4.18: Automatic Dialect Detection in Arabic Broadcast Media
Authors: Bilal Belainine and Fatiha Sadat
Country: Canada

Abstract: We describe our participation at the Arabic Multi-Genre Broadcast (MGB) Challenge for ASRU 2017. In this shared task, we used several supervised learning methods that emphasize only on labelled data, involving Voting Ensemble, Multi-Layer perceptron and Logistic classifiers, and their combinations, to discriminate and distinguish between four major Arabic dialects: Egyptian, Levantine, North African, Gulf and Modern Standard Arabic (MSA). Our best results were obtained during this shared task on closed submission using the Voting Ensemble with an overall accuracy of 56.10, followed by the simple logistic and Multi-Layer perceptron with an overall accuracy of 54.56 and 53.55, respectively.

Native language: وصفنا مشاركتنا في تحدي البث متعدد الأنواع العربية (MGB) لـ ASRU 2017. في هذه المهمة المشتركة ، استخدمنا العديد من أساليب التعلم التي تخضع للإشراف والتي تركز فقط على البيانات المصنفة ، بما في ذلك مجموعة التصويت ، المصنفات متعددة الطبقات والمصنفات اللوجستية ، ومجموعاتها ، للتمييز والتمييز بين أربع لهجات عربية رئيسية: مصر ، بلاد الشام ، شمال إفريقيا ، الخليج والعربية الحديثة المعاصرة (MSA). تم الحصول على أفضل النتائج خلال هذه المهمة المشتركة عند التقديم المغلق باستخدام فريق التصويت مع دقة إجمالية قدرها 56.10 ، تليها الإدراك البسيط لوجستي ومتعدد الطبقات مع الشاملة دقة 54.56 و 53.55 ، على التوالي.

P.4.19: The ADLaM Story
Authors: Abdoulaye Barry and Ibrahima Barry
Country: United States

Abstract: Thirty years ago, my brother and I created ADLaM to accurately transcribe all the sounds of the Fulfulde language spoken by more than 60 million people in Africa. The alphabet contains 28 letters and it was encoded in 2016 in Unicode 9.0. Since then our goal has been to try to get ADLaM working on every smartphone and computer. ADLaM is currently supported by Google through Android 8 and Chrome, and also by Microsoft through Windows 10. The script is now used from Gambia to Sudan by people who are very excited to use their language in their everyday life.

Native language: قصة منذ ثلاثين عامًا ، أسسني أنا وشقيقي ADLaM لنسخ جميع أصوات لغة الفولفولدي بدقة التي يتحدث بها أكثر من 60 مليون شخص في إفريقيا. يحتوي الأبجدية على 28 حرفًا وتم ترميزها في عام 2016 في Unicode 9.0. منذ ذلك الحين ، كان هدفنا هو محاولة الحصول على ADLaM للعمل على كل هاتف ذكي وكمبيوتر. ADLaM مدعوم حاليًا من Google من خلال Android 8 و Chrome ، وأيضًا من Microsoft من خلال Windows 10. يتم استخدام البرنامج النصي الآن من غامبيا إلى السودان من قبل أشخاص متحمسين جدًا لاستخدام لغتهم في حياتهم اليومية.

P.5.1: Linguistic Linked Open Data for All
Authors: John McCrae and Thierry Declerck
Country: Germany

Abstract: In this poster we show how to increase the uptake of language technologies for all by exploiting the combination of linked data and language technologies, that is Linguistic Linked Open Data (LLOD), to create ready-to-use multilingual data, also for low-resourced languages. The project Prêt-à-LLOD develops tools for the transformation and linking of datasets and apply these to both data and metadata in order to provide multi-portal access to heterogeneous data repositories. Prêt-à-LLOD implements a new methodology for building data value chains applicable to all types of language resources and language technologies that can be integrated by means of semantic technologies.

Native language: Nous montrons comment accroître l'adoption des technologies langagières pour tous en exploitant la combinaison de données liées (Linked Data) et de technologies langagières, c'est-à-dire l'infrastructure du Linguistic Linked Open Data (LLOD), pour créer des données multilingues prêtes à l'usage même pour les langues à faibles ressources. Le projet Prêt-à-LLOD développe des outils de transformation et de mise en relation des ensembles de données linguistiques et les applique à la fois aux données et aux métadonnées afin de fournir un accès multi-portail à des référentiels de données langagières hétérogènes. Prêt-à-LLOD met en œuvre une nouvelle méthodologie de construction de chaînes de valeur de données applicables à tous les types de ressources et de technologies langagières pouvant être intégrées par le biais des technologies sémantiques.

P.5.2: Bangla Text and Spoken Language Technology
Authors: Professor Dr. Mohammad Nurul Huda
Country: Bangladesh

Abstract: The poster will primarily focus on the following technologies developed by my team in NLP field. In addition, the major problems and obstacles faced in the development process will be addressed and analyzed. Some Developed NLP Tools are: Spell Checker, Question Answering System, Bangla TTS, Bangla to IPA Converter, Machine Translator, Bangla Sentiment Analyzer, Bangla-English Parallel Corpus. Moreover, in near future, we will develop Sign Language Tool and Screen Reader Tool for vision impaired people.

Native language: পোস্টারটি প্রাথমিকভাবে এনএলপি ক্ষেত্রে আমার দল দ্বারা নির্মিত নিম্নলিখিত প্রযুক্তিগুলিতে ফোকাস করবে। এছাড়াও, তৈরির প্রক্রিয়ায় যে বড় সমস্যা ও বাধার মুখোমুখি হয়েছিল সেগুলি সমাধান করে বিশ্লেষণ করা হবে। কয়েকটি তৈরিকৃত এনএলপি টুলস হল: বানান পরীক্ষক, প্রশ্ন উত্তর সিস্টেম, বাংলা টিটিএস, বাংলা থেকে আইপিএ রূপান্তরকারী, মেশিন অনুবাদক, বাংলা সেন্টিমেন্ট বিশ্লেষক, বাংলা-ইংলিশ প্যারালাল কর্পাস। তদুপরি, অদূর ভবিষ্যতে, আমরা দৃষ্টি প্রতিবন্ধী ব্যক্তিদের জন্য সাইন ল্যাঙ্গুয়েজ টুল এবং স্ক্রিন রিডার সরঞ্জাম তৈরি করব।

P.5.3: Envisioning a Trilingual Machine Translation System for the Language Pairs –<Tamang –English –Nepali>
Authors: Bal Krishna Bal, Amrit Yonjan Tamang and Lasang Jimba Tamang
Country: Nepal

Abstract: A Machine Translation system is a system that provides a gist or tentative translation in any target language for any given text in the source language. In the context of the degrading number of Tamang speakers among the younger generations of the Tamang community and in other contexts where the knowledge of English and/or Nepali may not necessarily be good among Tamang speakers, technology like MT system can help establish a strong stake or identity from a language empowerment or enabling perspective.

Native language: कुनै पनि यान्त्रिक अनुवाद प्रणालीले श्रोत भाषामा रहेको पाठलाई लक्षित भाषामा ठोस वा भावानुवाद गर्दछ। यस्तो प्रणाली दुई वा सोभन्दा बढी भाषी समुदायका बिच भाषिक तथा डिजिटल खाल्डो (समस्या) निवारण गर्न ठूलो भूमिका खेल्दछ। तामाङ समुदायको युवा पिंढीमाँझ आफ्नो तामाङ भाषा सिक्ने सन्दर्भमा देखिएको उदासिनताका साथसाथै कतिपय सन्दर्भमा अङ्ग्रेजी र नेपाली भाषाको ज्ञान उनीहरूमा कम भएको परिस्थितिलाई मध्यनजर राख्ने हो भने यान्त्रिक अनुवाद जस्तो प्रणालीले भाषाको पहिचान स्थापित गर्नका साथै भाषाको प्रबर्धनमा मद्दत प्रुयाउँन सक्छ। यस पोस्टरमा यस्तो त्रिभाषीय यान्त्रिक प्रणाली <तामाङ – अङ्ग्रेजी – नेपाली> को निर्माण सम्बन्धमा उच्च स्तरीय योजना प्रस्तुत गरिएको छ। उक्त योजनाअन्तर्गत भाषिक सामग्रीहरूको साथै न्युरल यान्त्रिक प्रणाली निर्माणबारे विस्तृत जानकारी समावेश छ। यस्तो प्रणालीको सफल विकासपश्चात समाजका विभिन्न तह तप्कामा प्रयोगमा ल्याई स्थानीय भाषा, संस्कृति र परम्पराको प्रबर्धन र विकास गर्न सकिने सम्भावना बोक्छ।

P.5.4: Keyman: High Fidelity Text Input for All Languages
Authors: Marc Durdin, Sok Makara, Joshua Horton and Ty Rasmey
Country: Australia

Abstract: While there are established conventions for typing Khmer using the Unicode Standard, existing systems provide little assistance to users in following the conventions which are thus often ignored. When typing Khmer text, users find that words can be constructed in multiple ways, all of which look 'correct' on-screen. Furthermore, some aspects of Khmer as implemented by common operating systems deviate from the Unicode Standard. This leads to a number of negative outcomes, including phishing and spoofing security risks, poor searchability and complications with natural language processing. Keyman provides a solution for these problems for Khmer and other languages.

Native language: ទោះបីជាមានគោលការណ៍សម្រាប់ការវាយអក្សរខ្មែរដោយប្រើ “យូនីកូដស្តង់ដារ” ក៏ដោយ ក៏ជំនួយដែលបានផ្តល់ឱ្យអ្នកប្រើប្រាស់ក្នុងការអនុវត្តតាមគោលការណ៍ទាំងនោះ នៅមានតិចតួចនៅឡើយ។ នេះជាហេតុនាំឱ្យអ្នកប្រើប្រាស់មិនសូវចាប់អារម្មណ៍អើពើ។ ពេលវាយអត្ថបទជាភាសាខ្មែរ អ្នកប្រើប្រាស់យល់ថាគេអាចវាយពាក្យបានច្រើនរបៀប ហើយវាមើលទៅត្រឹមត្រូវដូចគ្នានៅលើអេក្រង់។ ជាងនេះទៅទៀត ផ្នែកខ្លះនៃភាសាខ្មែរដែលត្រូវបានអនុវត្តដោយប្រព័ន្ធប្រតិបត្តិការពេញនិយម មានលក្ខណៈល្អៀងពីការកំណត់របស់ “យូនីកូដស្តង់ដារ”។ បញ្ហានេះនាំឱ្យមានលទ្ធផលអវិជ្ជមានជាច្រើន ដូចជា៖ ហានិភ័យផ្នែកសុវត្ថិភាពដោយសារការឆបោកនិងក្លែងបន្លំ លទ្ធភាពក្នុងការស្វែងរកមានកម្រិតទាប និង ភាពស្មុគស្មាញក្នុងដំណើរការភាសាបែបធម្មជាតិ។ ការស្រាវជ្រាវនេះរកឱ្យឃើញនូវបញ្ហានានាក្នុងការអ៊ិនកូដនៃភាសាខ្មែរដោយ “យូនីកូដស្តង់ដារ”។ ឃីមែន (Keyman) មានដំណោះស្រាយចំពោះបញ្ហាទាំងនេះ សម្រាប់ភាសាខ្មែរនិងភាសាផ្សេងទៀតផងដែរ។

P.5.5: Digital archiving and museum for language documentation and revitalization in Japan
Authors: Natsuko Nakagawa, Masahiro Yamada, Kenan Celik, Nobuko Kibe and Yukinori Takubo
Country: Japan

Abstract: There are eight (UNESCO), twelve (Ethnologue), or more (intelligibility) endangered languages/dialects in Japan. We present a database and digital archiving space that NINJAL (National Institute for Japanese Language and Linguistics) is developing for all of these languages where individual researchers or language communities can deposit their field data, language documentation, or audio-visual recordings. Two major features of the database/archiving space include (i) that it is a Japanese-mediated database/archive and thus virtually everyone in Japan can use it, and (ii) that it comes with an online exhibition space so that archiving is tightly connected to public use of the deposited items.

Native language: 日本には8（ユネスコ）、12（エスノローグ）、もしくはそれ以上（相互理解性）の消滅の危機に瀕した言語・方言が存在する。本発表は国立国語研究所がこれらの言語・方言のために開発中の、個別の研究者や言語コミュニティが利用可能なデータベースおよび電子的アーカイブスペースについて報告する。データベース・アーカイブスペースは以下の二つの特徴を持つ。(i) 日本語によるデータベース・アーカイブスペースであり、日本に住む誰もが利用可能である。(ii) オンライン展示スペースが付随し、アーカイブされるデータが社会一般に対する公開と密接に結びついている。

P.5.6: Project Mélange: Speech and Language Technologies for Code-switching
Authors: Sunayana Sitaram, Monojit Choudhury and Kalika Bali
Country: India

Abstract: Code-switching is the use of two or more languages in the same utterance or conversation, and is common in multilingual communities across the world. Project Mélange aims to process, understand and generate code-switched speech and text, so that technologies that interact with multilinguals can be natural and effective. In this poster, we present an overview of our research in the following areas 1. Data collection and generation 2. Core NLP and speech technologies (Language ID, Part of Speech tagging, Language Modeling, Speech Recognition and Synthesis) 3. sociolinguistics and pragmatics using Twitter data 4. user studies on dialogue and discourse

Native language: एक ही वार्तालाप में दो या दो से अधिक भाषाओं के उपयोग को कोड-स्विचिंग कहा जाता है, जो कि दुनिया भर के बहुभाषी समुदायों में आम है। प्रोजेक्ट मिलांज का उद्देश्य कोड-स्विच किए गए भाषण और पाठ को संसाधित करना, समझना और रचना करना है, ताकि बहुभाषियों के साथ बातचीत करने वाली प्रौद्योगिकियां प्राकृतिक और प्रभावी हो सकें। इस पोस्टर में हम निम्नलिखित क्षेत्रों में अपने शोध का एक अवलोकन प्रस्तुत करते हैं: 1. डेटा संग्रह और संश्लेषण 2. मौलिक भाषण प्रौद्योगिकियां (भाषा निर्धारण, शब्द के भेद निर्णय, भाषा मॉडलिंग, भाषण प्रतिलेखन और संश्लेषण) 3. सामाजिक और व्यावहारिक भाषाविज्ञान 4. संवाद और संभाषण पर उपयोगकर्ता अध्ययन।

P.5.7: Providing smart, open fonts for the world's language communities

Authors: Martin Raymond and Peter Martin

Country: United Kingdom

Abstract: SIL's Language Technology team has designed over thirty families of fonts, covering twenty different scripts, many of them Asian scripts, and supporting thousands of languages. All our fonts are released under the SIL Open Font License, enabling them to be distributed and modified freely. They also use font technologies, Graphite and OpenType, to handle complex writing systems.

The poster will display samples of several fonts, illustrating where in the world they are used, and listing some of the languages they support. Our Andika font was specifically designed for literacy use, and some of its graphic features will be illustrated.

Native language: L'équipe Language Technology de SIL a conçu plus de trente fontes (polices de caractères) dans vingt alphabets différents (beaucoup sont d'origine asiatique) pour assurer une bonne prise en charge dans plusieurs milliers de langues. Toutes nos fontes sont publiées sous « SIL Open Font License », ce qui permet à chacun de les diffuser et de les modifier librement. Elles tirent parti des technologies « Graphite » et « OpenType » pour gérer correctement plusieurs systèmes d'écriture complexes.

Le poster montrera une sélection de ces fontes, illustrera où elles sont actuellement utilisées à travers le monde et dans quelles langues on peut écrire grâce à elles. Andika a été tout spécialement conçue pour l'alphabétisation et certaines de ses caractéristiques graphiques seront aussi détaillées.

P.5.8: InaNLP: Indonesian Natural Language Processing Tools API

Authors: Ayu Purwarianti, Dessi Puji Lestari and Teguh Eko Budiarto

Country: Indonesia

Abstract: We've developed InaNLP, an Indonesian Natural Language Processing Tools API, which consists of several NLP tools that are easily integrated into a text processing module. InaNLP consists of lexical, syntactical and text classification modules, such as POS Tagger, named entity tagger, dependency parser, constituent parser, word normalizer, quotation extraction, document level and concept level sentiment analysis, and topic classification. These modules were built using deep learning algorithms with our own annotated data. The data annotation process was conducted by Indonesian linguists. In this poster, we will show the performance score of several InaNLP modules.

Native language: Kami mengembangkan InaNLP, API Kakas Pemrosesan Bahasa Alami Indonesia, yang terdiri dari beberapa kakas NLP yang mudah diintegrasikan ke dalam modul pemrosesan teks. InaNLP terdiri dari modul klasifikasi leksikal, sintaksis dan teks, seperti POS Tagger, entitas nama tagger , parser dependensi, parser konstituen, normalisasi kata, ekstraksi kutipan, analisis sentimen level dokumen, analisis sentimen level konsep, dan klasifikasi topik. Modul-modul ini dibangun menggunakan algoritma pembelajaran yang mendalam (deep learning) dengan data yang dianotasi sendiri. Proses anotasi data dilakukan oleh ahli bahasa Indonesia dan terdiri dari beberapa langkah seperti persiapan pedoman anotasi, pelabelan data dan pengecekan kualitas. Dalam poster ini, kami akan menunjukkan skor kinerja beberapa modul InaNLP.

P.5.9: The Pangloss Collection: an open archive of under-documented languages designed with Natural Language Processing in view

Authors: Séverine Guillaume, Balthazar Do Nascimento and Alexis MICHAUD

Country: France

Abstract: The Pangloss Collection was created by the research centre langues et civilisations à tradition orale (LACITO) in the 1990s, as a natural extension of traditional methods in linguistic fieldwork. As of 2019, the Pangloss Collection hosts about 170 languages, with 1900 hours of recordings (about 70% are transcribed and annotated). The resources in the Pangloss Collection benefit from long-term archiving services. Almost all resources are open access, so they are available for a variety of uses, for specialists but also for the general public and, last but not least for research in Natural Language Processing.

Native language: La Collection Pangloss : une archive ouverte de langues peu documentées conçue pour faciliter des emplois en Traitement Automatique des Langues

La Collection Pangloss a été créée par le laboratoire de langues et civilisations à tradition orale (LACITO), dans les années 90, dans le prolongement des méthodes classiques d'enquête et d'analyse de la linguistique de terrain. En 2019, la Collection Pangloss regroupe environ 170 langues, avec 1900h d'enregistrements (dont environ 70% transcrit et annoté). Les ressources de la Collection Pangloss bénéficient de services d'archivage pérenne. La quasi-totalité des ressources est en accès libre, elles sont donc disponibles pour divers usages : découverte, enseignement, recherche. Recherche en linguistique et anthropologie mais aussi, grâce au numérique, recherche dans le traitement automatique des langues.

P.5.10: Multi-lingual Support in Connective Learning Scheme for Refining and Connecting the Open Educational Videos
Authors: Virach Sornlertlamvanich, Nannam Aksorn and Thatsanee Charoenporn
Country: Japan

Abstract: Tons of educational videos are available online. It is a big burden for learners to figure out the videos they need in the preferred time and language. Not all videos are suitable for learning according the length and presentation components. According to the Sweller's cognitive load theory, the working memory in learning process is very limited, the learner must be selective to what information from sensory memory to pay attention. In the connective learning, we effectively apply NLP approach to refine the video subtitle in archiving, translating, summarizing, classifying, and labelling the relevant keywords to create the multi-lingual learner-friendly environment.

Native language: ปัจจุบันมีวิดีโอเพื่อการศึกษาที่เผยแพร่ออนไลน์มากมาย จึงเป็นการไม่สะดวกสำหรับผู้เรียนในการหาวิดีโอที่ต้องการได้ ซึ่งส่วนใหญ่ก็ต้องเลือกดูบางส่วนก่อนเพื่อให้ทราบเนื้อหา และวิดีโอส่วนใหญ่ก็เป็นภาษาอังกฤษหรือภาษาอื่นๆ ที่ผู้เรียนไม่สันทัดมากนัก จากทฤษฎีการเรียนรู้ (cognitive load theory) ของ Sweller ที่ได้กล่าวไว้ว่าในกระบวนการเรียนรู้ผู้เรียนจำเป็นต้องอาศัยหน่วยความจำชั่วคราว (working memory) ซึ่งมีพื้นที่จำกัด ดังนั้นเพื่อให้การเรียนรู้มีประสิทธิภาพสูงสุด งานวิจัยนี้ได้นำเสนอการใช้การประมวลผลภาษาธรรมชาติเพื่อช่วยในการจัดเก็บคำบรรยายประกอบ แปลคำบรรยาย ย่อความ จำแนก และสกัดคำสำคัญสำหรับการนำเสนอบทเรียนด้วยภาษาที่ต้องการและปรับแต่งให้เป็นวิดีโอที่เหมาะสมตามทฤษฎีการเรียนรู้

P.5.11: Promoting and Preserving Philippine Culture and Languages through Language Technologies
Authors: Ethel Ong, Nathalie Rose Lim-Cheng, Charibeth Cheng and Edward Tighe
Country: Philippines

Abstract: Advances in language technologies enabled the computational representation, processing and generation of human languages that gave computers the ability to analyze varying text and participate in human conversations. Digital resources such as lexicons and textual corpora led to the development of intelligent agents that can perform tasks in language translation and generation, sentiment analysis, fake news detection, and text mining. In this poster, we present our work in preserving and promoting the Philippine culture and language through computer-generated stories and descriptions of museum artifacts, and the analysis of bilingual social media posts to detect public sentiments and monitor public health.

Native language: Ang pagsulong sa teknolohiyang pang-wika na nagsasagawa ng representasyon, pagproseso at automatikong pagsulat ang nagbigay sa kompyuter ng kakayahang pag-aralan ang iba't ibang teksto at makilahok sa pakikipag-usap sa tao. Ang mga yamang wika tulad ng talasalitaan at mga sulatin ang nagpalaganap sa Artificial Intelligence upang magsagawa ng pagsalin at pagsulat ng wika, pagsuri ng sentimento, pagtuklas ng huwad na balita, at pagmina ng teksto. Nais naming ipakita sa poster na ito ang iba't ibang mga gawain sa pagpapanatili at pagtataguyod ng kultura at wika ng Pilipinas sa pamamagitan ng mga kwento na nilikha ng kompyuter at mga paglalarawan ng mga bagay sa museo, at pagsusuri ng mga sulatin ng publiko sa social media upang makita ang mga damdamin ng publiko at subaybayan ang kalusugan ng publiko.

P.5.12: Improvement of Thai NER and the Corpus
Authors: Thatsanee Charoenporn and Virach Sornlertlamvanich
Country: Japan

Abstract: Thai named entity (NE) corpus is rarely found though the named entity recognition (NER) task can make a big contribution in processing the huge amount of available texts. We propose an iterative NER refinement method using BiLSTM-CNN-CRF model with word, part-of-speech, and character cluster embedding to clean up the existing NE tagged corpus due to its inconsistent and disjointed annotation. As a result, in the newly generated corpus, we obtain 639,335 NE tags, much larger than the original size of 172,232 NE tags. The generated model by the newly generated corpus also improves the NER F1-score 16.21% to mark 89.22%.

Native language: การพัฒนาคลังข้อความภาษาไทยสำหรับการประมวลผลภาษาธรรมชาตินั้น มีประเภทและปริมาณเพิ่มมากขึ้น แต่คลังข้อความชื่อเฉพาะภาษาไทย หรือ Thai Name Entity Corpus ยังคงมีทั้งจำนวนที่จำกัด แม้ว่างานวิจัยด้านการรู้จำชื่อเฉพาะ (Name Entity Recognition: NER) จะส่งผลต่อความถูกต้องของการประมวลผลข้อความเป็นอย่างมากก็ตาม งานวิจัยนี้ เสนอวิธีการปรับแต่ง NER แบบวนซ้ำโดยใช้แบบจำลอง BiLSTM-CNN-CRF ประกอบกับคำแวดล้อม หน้าที่ของคำ และกลุ่มอักขระข้างเคียง เพื่อปรับปรุงคลังข้อความชื่อเฉพาะภาษาไทย จากเดิมจำนวน 172,232 ชื่อ ให้มีความถูกต้อง แม่นยำ และสอดคล้องกัน ผลการวิจัยพบว่า คลังข้อความชื่อเฉพาะภาษาไทยที่ปรับปรุงขึ้น ประกอบด้วยคำและป้ายระบุชื่อเฉพาะ (Tags) จำนวนถึง 639,335 ชื่อ ทั้งนี้ ผลการปรับปรุงคลังข้อความชื่อเฉพาะด้วยแบบจำลองที่นำเสนอนี้สามารถกำกับชื่อเฉพาะภาษาไทยมีความถูกต้อง ที่วัดด้วยค่า F1-score ได้ที่ 89.22 เปอร์เซ็นต์ ซึ่งให้ผลที่ดีกว่าแบบจำลองที่สร้างด้วยคลังข้อความเดิมถึง 16.21 เปอร์เซ็นต์

P.5.13: Deploying Language Technologies for Underserved Communities
Authors: Kalika Bali, Monojit Choudhury, Sunayana Sitaram and Sebastin Santy
Country: India

Abstract: Gondi, a South-Central Dravidian language in the vulnerable category on UNESCO's Atlas of the Worlds Languages in Danger (Moseley,2010). Spoken by nearly 3 million people (India, 2011) it is also one of the least resourced languages in India, with little available data and technology. In this poster, we will present Adivasi Radio, a mobile application that provides the Gond tribal community with access to information. We believe that by focusing on Gondi, we will not only empower the Gondi community but also help understand and create a framework to serve as a guide for introducing language technologies in under-served communities.

Native language: गोंडी, यूनेस्को के एटलस ऑफ़ द वर्ल्ड्स लैंग्वेजेस ऑफ़ डेंजर (मोसेली, 2010) में एक दक्षिण-मध्य द्रविड़ भाषा की असुरक्षित श्रेणी में है। लगभग ३0 लाख लोगों (भारत जनगणना, २०११) द्वारा बोली जाने वाली यह भारत की सबसे कम पुनर्जीवित भाषाओं में से एक है, जिसमें बहुत कम डेटा और तकनीक उपलब्ध है। इस पोस्टर में, हम आदिवासी रेडियो पेश करेंगे, जो एक मोबाइल एप्लिकेशन है जो गोंड आदिवासी समुदाय को सूचना तक पहुंच प्रदान करती है। हमारा मानना है कि गोंडी पर ध्यान केंद्रित करने से, हम न केवल गोंडी समुदाय को सशक्त बनाएंगे, बल्कि पिछड़े समुदायों में भाषा प्रौद्योगिकी को शुरू करने के लिए एक मार्गदर्शक के रूप में काम करने में मदद करेंगे।

P.5.14: Language Technologies at the University of the Philippines DIliman
Authors: Angelina Aquino and Rhandley Cajote
Country: Philippines

Abstract: Research initiatives on language technologies at the University of the Philippines Diliman are the result of collaborations between the disciplines of engineering and linguistics. These efforts can be categorized into database development, speech processing, text processing, and linguistic analysis. We highlight some of the recent developments in speech processing, speech synthesis, and natural language understanding for the Filipino language and other major languages in the Philippines. We demonstrate some of these technologies in applications such as closed-captioning, automated reading tutors, and literacy assessment. We also discuss efforts towards documentation, preservation, and historical linguistics for indigenous Philippine languages.

Native language: Ang mga pananaliksik sa pangwikang teknolohiya sa Unibersidad ng Pilipinas Diliman ay bunsod ng pakikipagtulungan ng mga disiplina ng inhinyeriya at linggwistiks. Ang mga pagsisikap na ito ay maaaring maiuri sa pagbuo ng mga korpus, pagproseso ng pagsasalita, pagproseso ng teksto, at pagsusuri ng wika. Binibigyang-diin namin ang ilan sa mga kamakailang pag-unlad sa pagproseso ng pagsasalita, sintesis ng wika, at pag-unawa ng wikang Filipino at iba pang mga pangunahing wika sa Pilipinas. Ipinapakita namin ang ilan sa mga gamit ng teknolohiyang ito tulad ng closed-captioning, gabay sa pagbasa, at pagsusuri ng kasanayan sa pagbasa. Itinatatalakay din namin ang mga gawain tungo sa dokumentasyon, pagpapanatili, at pangkasaysayang linggwistiks para sa mga katutubong wika ng Pilipinas.

P.5.15: Languages and Technology in Bhutan
Authors: Tenzin Namgyel
Country: Bhutan

Abstract: Dzongkha is the national language of Bhutan. Officially, there are 19 languages in Bhutan according to the survey carried out in 1991. Dzongkha efficiently serves as the official language of Bhutan while Chökê serves as the language of Dharma and liturgy; and English is apparently used as the necessary foreign language while mother tongues are used at the grass-roots level. Though we are working on development of technology for Bhutanese languages, we are acutely challenged with the lack of expertise and funding.

Native language: རྫོང་ཁ་འདི་འབྲུག་གི་རྒྱལ་ཡོངས་སྐད་ཡིག་ཨིན། སྡེབ་ཚ་1991 ལུ་འབད་མི་ཞིབ་འཚུགས་དང་འཁྲིལ་ཡས་བ། གཞུང་འབྲེལ་བ། འབྲུག་ལུ་སྐད་ཡིག་19 ཡོད། རྫོང་ཁ་འདི་གཞུང་འབྲེལ་ཀྱི་ལས་གཞི་སྐད་ཡིག་ཨིན་དང་། གཡུས་སྐད་གཞན་རྟེན་སྡུག་ཅིང་ཁྲིམ་རྟའི་ཆུ་ཤེལ་ཁའི་གཡུག་ཁང་ཐ་རྩོང་རྩང་སྐབས་ནི་ཆོས་དང་ལྷས་བྱེད་ཡོད་ཨིན། དེ་འཐས་ ཕིང་སྐད་འདི་མེད་ཐབས་མེད་པའི་ཕྱི་རྒྱལ་ཀྱི་ཀུན་སྐད་འདི་རོམ་ཀྱི་དོན་ལུ་ལག་ལེན་འཐབ་དོ་ཡོད་པ་ཨིན། ང་བཅས་ཀྱིས་འབྲུག་པའི་ཁ་སྐད་ཀྱི་འཕྲུལ་རིག་གོང་འཕེལ་གཏང་རའི་ལུ་རྩ་འབད་དོ་ཡོད་པ་ཨིན་རུང་ ལྟོག་གས་གྱུབ་ཅན་ཀྱི་མི་རྟོལ་བང་མ་བདུལ་རྩ་མེད་པའི་དཀའ་ངལ་སྤྲོལ་ར་འབྲུང་དེ་ཡོད་པ་ཨིན།

P.5.16: Language technology at MILE Lab, Indian Institute of Science
Authors: Ramakrishnan AngaraiGanesan
Country: India

Abstract: Medical Intelligence and Language Engineering Laboratory at the Department of Electrical Engineering, Indian Institute of Science is engaged in research, development and deployment of technologies that facilitate individuals, organisations and governments to easily carry out their daily operations, amidst the existence of multiple, living Indian languages. Using the OCR and TTS developed, hundreds of books have been converted to Braille and audio books, which are being used by blind students. Also developed are online handwriting recognition, browser-based tools to instantly convert text across different Indian scripts that enable people to read text in other languages through their own script.

Native language: இந்தியாவில் பெங்களூரிலுள்ள இந்திய அறிவியல் பயிலகத்தில் மின் பொறியியல் துறையிலுள்ளது மருத்துவ நுண்ணறிவு மற்றும் மொழிப் பொறியியல் ஆய்வகம், இந்த ஆய்வகம், மொழித் தொழில்நுட்பங்களின் ஆராய்ச்சி, உருவாக்கம் மற்றும் பயனாக்கம் ஆகியவற்றில் ஈடுபட்டுள்ளது. இதனால், பல்வேறு வழங்கு மொழிகள் இந்தியாவில் இருப்பினும், தனிநபர்கள், நிறுவனங்கள் மற்றும் அரசாங்கங்கள் தங்கள் அன்றாட நடவடிக்கைகளை எளிதில் மேற்கொள்ள இயலும்.

இங்கு உருவாக்கப்பட்ட அச்சிடப்பட்ட உரை எழுத்து உணர்வி மற்றும் பேச்சு உருவாக்கி ஆகிய மென்பொருட்களைப் பயன்படுத்தி, அறுநூற்றுக்கும் மேற்பட்ட பள்ளி, மற்றும் கல்லூரிப் புத்தகங்கள் பிரெய்ல் மற்றும் பேசும் புத்தகங்களாக மாற்றப்பட்டுள்ளன. அவை பார்வையற்ற மாணவர்களால் பயன்படுத்தப்படுகின்றன. டேப்லெட்களில் எழுதப்பட்ட உரையை உணரும் கருவியும் உருவாக்கப்பட்டுள்ளது. மற்றும் வெவ்வேறு இந்திய லிபிகளிலுள்ள உரையை உடனடியாக பிற மொழிகளில் மாற்றுவதற்கான உலாவி அடிப்படையிலான மென்பொருளால் தங்கள் சொந்த எழுத்துரு மூலமே படிக்க முடியும்.

P.5.17: MULTILINGUAL PROFILE OF INDIA
Authors: Udaya Narayana Singh
Country: India

Abstract: South Asia is the largest repository of multi-lingual and pluri-ethnic civilization, dating back to 2600 BC. Beginning from the Indus Valley Civilization, traces of which were found in India, Pakistan, and Afghanistan, it had two writing systems - Brāhmī (5th CBC) & Kharoṣṭhī (4thCBC) both widely used in many Asian countries. With its 1,576 rationalized mother-tongues and 1,796 other mother-tongues, India has 122 major languages and 200 minor ones. Here, 96% speak only 20-odd IA & Dravidian languages included in the 8th Schedule of the Constitution, and 50 were found to have a high vitality and use.

Native language: দক্ষিণ এশিয়া হলো বিশ্বের বহুভাষিক ও বহুসাংস্কৃতিক সভ্যতা গুলির মধ্যে একটি অন্যতম শ্রেষ্ঠ উদাহরণ. খ্রিস্ট পূর্ব ২৬০০ সালে এর ভাষিক মিলন ও মিশ্রণ প্রক্রিয়া শুরু হয়েছিল. সিন্ধু সভ্যতার সময় থেকেই এর বহু উদাহরণ আধুনিক ভারত, পাকিস্থান ও আফগানিস্থানে পাওয়া যায়. লিখিত রূপে উত্তরণ দুটি লিপির ব্যবহার দিয়ে শুরু হয়. একটি হলো 'ব্রাহ্মী' লিপি (খ্রিস্ট পূর্ব পঞ্চম শতক থেকে), অন্যটি খরোষ্ঠী (খ্রিস্ট পূর্ব চতুর্থ শতক). এই দুটিই এশিয়ার বহু দেশে ছড়িয়ে পড়েছিল. এখন প্রায় ১৫৭৬ নথিভুক্ত মাতৃভাষা ও ১৭৯৬ অন্য মাতৃভাষা সমেত ১২২ টি প্রমুখ ভাষা আছে ভারতে যার অতিরিক্ত রয়েছে শ' দুয়েক ছোট ছোট ভাষা সম্প্রদায়. শতকরা ৯৬ জন এর মধ্যে কুড়িটি ভারোপিয় ও দ্রাবিড় ভাষায় কথা বলেন যেগুলি ভারতীয়সংবিধানের অষ্টম অনুসুচিতে স্থান পেয়েছে. এগুলির মধ্যে পঞ্চাশটি ভাষা এমন যাতে প্রচুর লেখা ও ছাপা হচ্ছে.

P.5.18: Mediating Multilingualism
Authors: Udaya Narayana Singh, Esha Jainiti, Rusha Mudgal and Anwita Maiti
Country: India

Abstract: "Mediating Multilingualism" is a collaborative project for language revival and survival between Amity University Haryana and the University of Highlands & Islands, UK. It encourages video ethnographic documentation of some marginalized speech communities, with special focus on the Indian context based on work done earlier in Scotland and Ireland. Colloquiums, seminars and workshops organized focused on protecting and promoting the endangered languages of India. The primary aim is building capacity and networking with other academic partners, take up interdisciplinary research from both sociolinguistic and Communication technology angle. It will be followed by a desk-based monitoring, assessment and theoretical work.

Native language: “মেডিয়েটিং মাল্টিলিঙ্গুয়ালিজম” অ্যামিটি বিশ্ববিদ্যালয় হরিয়ানা (AUH) এবং ইউনিভার্সিটি অফ হাইল্যান্ডস এবং আইল্যান্ডস (UHI), স্কটল্যান্ড, ইউ-কে--র মধ্যে ভাষা পুনরুজ্জীবন এবং ভাষা সংরক্ষণ বিষয়ক আকাদেমিক মধ্যস্থতা করার সম্ভাবনা খুঁটিয়ে দেখার জন্য একটি সহযোগী প্রকল্প - যে ফোরামটি কিছু প্রান্তিক ভাষা ও সংস্কৃতি বিষয়ে সেই সব সম্প্রদায়ের আলোচনা এবং ভিডিও-র মাধ্যমে নৃতাত্ত্বিক ডকুমেন্টেশনকে উৎসাহিত করবে, এবং এর আগে স্কটল্যান্ড এবং আয়ারল্যান্ডে করা কাজের ভিত্তিতে ভারতীয় প্রসঙ্গে সেই অভিজ্ঞতা কতটা কাজে লাগানো যায় সেবিষয়ে বিশেষ মনোযোগ দেবে. কথোপকথন, সেমিনার এবং ওয়ার্কশপগুলি ভারতের আদিবাসী ও বিপন্ন ভাষা সুরক্ষিত ও প্রচারের দিকে দৃষ্টি নিবদ্ধ করবে । প্রাথমিক লক্ষ্য দক্ষতা তৈরি করা এবং অন্যান্য একাডেমিক অংশীদারদের সাথে নেটওয়ার্কিং করা, অন্তর্বিভাগীয় গবেষণা কার্যক্রম গ্রহণ করা, ভাষাতত্ব ও যোগাযোগ প্রযুক্তি দুই দিক থেকেই দক্ষতা অর্জন করা। এটি একটি পর্যবেক্ষণ, মূল্যায়ন এবং তাত্ত্বিক কাজ অনুসরণ করবে।

P.5.19: Technology Development for Indian Languages
Authors: Vijay Kumar and Dr S K Srivastava
Country: India

Abstract: Technology Development for Indian Languages (TDIL) Programme of Government of India has been sponsoring projects for development of Linguistic Resources, Standards and Technologies like Fonts, Unicode Typing Tool, Localized Open Source Software, Machine Translation Systems, Speech Technologies (TTS, ASR), Optical Character Recognition, etc. The developed prototypes are accessible through http://www.tdil-dc.in. A new initiative viz. “Natural Language Translation Mission” is being started with an objective to build a speech to speech translation system for major Indian languages in the domains like Science & Technology, Education, Healthcare, Law& Justice and Governance.

Native language: भारत सरकार का भारतीय भाषाओं के लिए प्रौद्योगिकी विकास (टीडीआईएल) कार्यक्रम भाषाई संसाधनों के विकास , मानकों और प्रौद्योगिकी जैसे फोंट, यूनिकोड टाइपिंग टूल, स्थानीयकृत ओपन सोर्स सॉफ्टवेयर, मशीन ट्रांसलेशन सिस्टम्स, स्पीच टेक्नोलॉजीज (टीटीएस, एएसआर), ओसीआर आदि के विकास के लिए परियोजनाओं को प्रायोजित कर रही है। विकसित किए गए प्रोटोटाइप को http://www.tdil-dc.in के माध्यम से देखा जा सकता है। विज्ञान और प्रौद्योगिकी, शिक्षा, स्वास्थ्य, कानून और न्याय , शासन जैसे क्षेत्रों में प्रमुख भारतीय भाषाओं के लिए स्पीच टु स्पीच ट्रांसलेशन सिस्टम का निर्माण करने के उद्देश्य से एक नई पहल की जा रही है ।

P.5.20: CREATING ACCESS TO OPENLY LICENSED EARLY READING RESOURCES IN ASIA'S INDIGENOUS LANGUAGES
Authors: Purvi Shah
Country: India

Abstract: StoryWeaver is an open-access digital platform with 17,000+ children’s storybooks in 200+ languages. 78 of these are Asian, including 16 Indigenous languages. The platform contains tools to read, translate, publish, download, print books – all for free. This poster presentation highlights how StoryWeaver uses technology and fosters communities to build a large repository of local language reading resources: from creating bilingual books to aid a child’s transition to the language of instruction in school, to facilitating the self-expression of Indigenous groups by publishing books in languages without writing systems, and finally creating sustainable self-publishing models by leveraging open licences.

Native language: स्टोरीवीवर बाल कहानियों का खुला मंच है जहाँ 200 से अधिक भाषाओं में 16000 से अधिक कहानियाँ मौजूद हैं। इनमें 78 एशियायी और 16 भारतीय भाषाएँ शामिल हैं। यहाँ कहानियाँ पढ़ी, अनुवाद, प्रकाशित, डाउनलोड और मुद्रित की जा सकती हैं- बिल्कुल निशुल्क। यह प्रस्तुति बताती है कि स्थानीय भाषाओं में पठन संसाधनों का विकास करने में स्टोरीवीवर किस प्रकार सहायक बना- द्विभाषी पुस्तकों के द्वारा स्कूली और मातृभाषाओं के बीच पुल बनके, क्षेत्रीय समूहों द्वारा अपनी बिना लिपि वाली भाषाओं में स्वाभिव्यक्ति का साधन बन, अंततः ओपन लाइसेंस का लाभ उठाकर कारगर स्वप्रकाशन मॉडल तैयार करके।

P.5.21: Conversational Bot for Eyesight Testing Automation
Authors: Ari Yanase, Thatsanee Charoenporn and Virach Sornlertlamvanich
Country: Japan

Abstract: The people’s visual acuity can be examined using a standard Snellen eye chart by determining a relationship between the sizes of certain letters viewed at certain distances, or a broken wheel vision chart. In general, we examine our eyesight by reading the letter “A” with 88 mm in height at 20 feet or 6 meters in distance. To lessen the burden of the ophthalmologist, we equip an NLP based chatbot with basic eyesight testing knowledge in a mobile app. With the natural conversation, the bot recognizes the response from the subject, and returns the prescriptions for eyeglasses measured in diopters.

Native language: 人々の視力測定には、文字のサイズなどを元に作成されたSnellenのモデルと壊れたタイヤにまつわるモデルが使われています。一般的には、６メートルもしくは２０フィート離れた距離から、高さ８８ミリで書かれたＡの文字が読めれば、１.０の視力を持っている、と言われています。私達が視力測定の基本的知識を備えた自然言語処理を元に作成されたチャットボットを備える、作成することで、眼科医の負担になる仕事を減らしました。また、自然言語を用いて、眼鏡の度を予測したりもできます。

P.5.22: Dictionary 4.0: Alternative Presentations for Indonesian Multilingual Dictionaries
Authors: Arbi Haza Nasution and Totok Suhardijanto
Country: Indonesia

Abstract: Building a multilingual dictionary for 719 languages in Indonesia is a challenging task. We have developed application to create the Leipzig-Jakarta list database for all indigenous languages in Indonesia. The database can be used to generate lexical similarity or lexical distance matrix between languages by comparing the word list. For starter, we covered 11 languages: Indonesian, Javanese, Sundanese, Madurese, Bima, Ternate, Tidore, Palembang Malay, Mandailing Batak, Malay, and Minangkabau. The application has two main features: exploring the existing translations and adding translations to a new language or editing existing translations through crowdsourcing. User acceptance test showed 3.48 / 4 score.

Native language: Membangun kamus multibahasa untuk 719 bahasa di Indonesia adalah tugas yang berat. Kami telah mengembangkan aplikasi untuk membuat pangkalan data daftar Leipzig-Jakarta untuk semua bahasa daerah di Indonesia. Pangkalan data tersebut dapat digunakan untuk menghasilkan kesamaan leksikal atau matriks jarak leksikal antar bahasa dengan membandingkan daftar kata tersebut. Sebagai permulaan, aplikasi ini mencakup 11 bahasa: Indonesia, Jawa, Sunda, Madura, Bima, Ternate, Tidore, Melayu Palembang, Batak Mandailing, Melayu, dan Minangkabau. Aplikasi ini memiliki dua fitur utama: menjelajahi terjemahan yang ada dan menambahkan terjemahan ke bahasa baru atau mengedit terjemahan yang ada melalui mekanisme urun daya dari masyarakat. Uji keberterimaan pengguna menunjukkan skor 3,48 / 4.

P.5.23: Speech Technology in three tonal languages of North-East India
Authors: Viyazonuo Terhiija, Samudra Vijaya and Priyankoo Sarmah
Country: India

Abstract: An account of speech systems implemented in three tonal languages of North East India is presented here. Angami, Ao, and Mizo are under-resourced languages of the Tibeto-Burman language family, spoken in North-East of India. The distribution of tones across the three languages vary from three to four tones. Automatic Speech Recognition systems were developed for Angami and Mizo languages using Kaldi toolkit. The performances of various versions of the system using different types of acoustic models (GMM-HMM, SGMM and DNN) are reported. In Ao, a dialect identification system was built that distinguishes two Ao dialects, namely, Changki and Mongsen.

Native language: Diepu kesezho nu North East India nu chiiwaketa die pfhephra se pie kepushiizhie. Angami, Ao mu Mizo seyakezha dieko Tibeto - Burman die kikru nu ba. Die hako puo pfhephra kezakeshii ki pfhephra se mu dia mese kekrei ba. Automatic speech recognition systems-e Angami mu Mizo die la Kaldi Toolkit se di chii kehielie. Die puo zho kekreikecii mhathokoe acoustic models (GMM-HMM, SGMM and DNN) kekreikecii se di chiikelie silie. Ao mia-e die si kekreilie kevi zho puo chiilie di uko die kenie Changki mu Mongsen si kekreilie kevi chiilie.

P.5.24: Bringing Zero-resourced Languages of Myanmar to the Digital World
Authors: Win Pa Pa
Country: Myanmar

Abstract: Accessing Technology in one's own language is promoting personnel, social and economic life of that one and also preserving the language. Myanmar language(Burmese) is one of low-resourced language although it is spoken by 33 millions people as their first language. Data scarcity of Myanmar language is a big challenge for language processing. Recent technologies and resources of Machine Translation, Automatic Speech Recognition and Speech synthesis are presented in the poster to promote Myanmar languages and its dialects' language processing.

Native language: နည်းပညာအသုံးပြုမှုကို မိမိတို့ ကိုယ်ပိုင် ဘာသာစကားဖြင့် ပြုလုပ်နိုင်ခြင်းသည် မိမိတို့၏ ကိုယ်ပိုင်ဘဝ၊ လူမှုပတ်ဝန်းကျင် နှင့် စီးပွားရေး တိုးကို မြင့်တင်ပေးရာ ရောက်ပြီး ထို ကိုယ်ပိုင်စကားကို ထိန်းသိမ်းရာလည်း ရောက်ပေသည်။ မြန်မာဘာသာစကားသည် မိခင် ဘာသာစကား အဖြစ် အသုံးပြုသူ ဒသရန်း ရိသော်လည်း အရင်းအမြစ် နည်းပါးသော ဘာသာစကားများတွင် တစ်ခု အပါအဝင်ဖြစ်သည်။ မြန်မာဘာသာ အချက်အလက် ရှာပါခြင်း သည် ဘာသာစကားနည်းပညာ အတွက် ကြီမားသော စိန်ခေါ်မှု တစ်ခုဖြစ်သည်။ ပိုစတာတွင် မြန်မာဘာသာစကားအတွက် ကက်မြင့်ဘာသာပြန်ခြင်း၊ အသံမှ စာသား သို့ပြောင်းခြင်း နှင့် စာသားမှ အသံသို့ပြောင်းခြင်း လုပ်ငန်း၏ လက်ရှိ နည်းပညာများ ကို ဖော်ပြထားပြီး မြန်မာ နှင့် တိုင်းရင်းသား ဘာသာစကားများ နှင့် ဒေသိယ စကားများ အတွက် နည်းပညာများ ကို လူအများ စိတ်ဝင်စားမှု မြင့်တင်ရန် ရည်ရွယ်ပါသည်။

P.5.25: Building Corpora for Under-Resourced Languages in Indonesia
Authors: Totok Suhardijanto and Arawinda Dinakaramani
Country: Indonesia

Abstract: Indonesia has known as the second most linguistically-diverse country, but ironically also known as a country with many under-resourced languages. In this poster, we present our attempt to develop language resources in Indonesian indigenous languages for linguistic research purposes. For the first stage, we developed corpora for Javanese, Sundanese, Malay/Indonesian, and Minangkabau which are chosen because of the number of speakers. This poster discusses the drawbacks and opportunities in our attempt to build Indonesian language corpora that are publicly accessible. The corpus application is still under development, but it is a good step to start compiling language corpora in Indonesia.

Native language: Indonesia dikenal sebagai negara paling kaya kedua dalam hal bahasa, namun ironisnya juga dikenal dengan negara paling sedikit sumber daya bahasanya. Poster ini menyajikan upaya kami dalam membangun korpora untuk bahasa-bahasa di Indonesia untuk kepentingan kajian bahasa. Pada tahap awal, dipilih bahasa Jawa, Sunda, Melayu/Indonesia, dan Minangkabau sebagai konten korpora karena bahasa-bahasa itu mempunyai jumlah penutur banyak. Poster ini akan membahas kendala dan kesempatan dalam menyusun korpora bahasa-bahasa di Indonesia. Aplikasi korpusnya pun masih dalam tahap pengembangan, namun ini merupakan langkah baik untuk mengembangkan korpus untuk bahasa-bahasa di Indonesia.

P.5.26: Language Resources and Technology Development Efforts for some Lesser-known Indian Languages
Authors: Ritesh Kumar, bornini lahiri, Atul Kr. Ojha, Mayank Jain and Deepak Alok
Country: India

Abstract: For the last few years, we have been involved in the development of language technologies and resources for some of the lesser-known Indian langauges viz. Magahi, Bhojpuri, Awadhi and Braj Bhasha. These languages (among others) have been largely marginalised and ignored and have low prestige and negative attitude towards them because of these being considered ‘illiterate’ and ‘rural’ varieties of Hindi. Our poster will showcase different kinds of corpora as well as basic technologies like part-of-speech tagger, morphological analyser as well as some applications like machine translation systems that have been developed for these languages.

Native language: पिछला कै साल से, हमनी मगही, भोजपुरी, अवधी आउ ब्रज भासा जईसन भासा, जेकरा पर बहुते कम काम होल हई, ओकर प्रौद्योगिकि आउ संसाधन के विकास में लगल ही । इ सब भासा (औ अइसन बड़ीमनी) के नजर अंदाज और हासिया पर कर देल गेल हे आउ कम परतिस्ठा आउ नीचा दृस्टि से देखल जा हे काहे की इ सब हिंदी भासा के देहाती आउ असीछित बोली के प्रकार समझल जा हे। हमनी के पोस्टर बिभिन्न प्रकार के कॉर्पोरा के साथ-साथ बुनयादी प्रौद्योगिकि जैसे की सब्द भेद टैगर (पार्ट ऑफ़ स्पीच टैगर), रूप सम्बन्धी बिश्लेसक (मोरफोलॉजिकल अनलाइज़र) आउ कुछ अनुप्रयोग सॉफ्टवेयर जैसे की मसीनी अनुवाद पद्धति/सिस्टम जे इसब भासा ला विकिसित कइल गइल हे ओकरा परदरसीत करत।

P.5.27: A 1000-language Collaborative Universal Dictionary and Universal Translator

Authors: David Yarowsky, Arya D. McCarthy, Garrett Nicolai, Winston Wu, Aaron Mueller, Dylan Lewis, Yingqi Ding, Abhinav Nigam, Emre Ozgu, Debanik Purkayastha, James Scharf and Kenneth Zheng

Country: United States

Abstract: We present JHU's Universal Dictionary and Universal Translator, covering 1000+ world languages in a broadly-accessible Android/IOS mobile phone and web-browser app, with 1,000,000+ planet-wide language pairs and 100's of under-resourced languages which have never had access to a substantial dictionary or machine translation capability. In addition to providing immediate access to a base vocabulary of 1500-20000 core vocabulary lemmas in all 1000+ languages, this novel app actively engages its users to contribute collaboratively to the universal dictionary in an easy-to-use and efficient way, with automatic suggestion of possible translations based on sound-shift transductions from related languages and pan-linguistic compositional constructions.

Native language: Presentamos el Diccionario Universal y el Traductor Universal de JHU, que cubren más de 1000 idiomas del mundo en una aplicación de navegador web y teléfono móvil Android / IOS de amplio acceso, con más de 1,000,000 de pares de idiomas en todo el planeta y cientos de idiomas con recursos insuficientes que nunca han tenido acceso a un diccionario sustancial o capacidad de traducción automática. Además de proporcionar acceso inmediato a un vocabulario base de lemas de vocabulario básico de 1500-20000 en los más de 1000 idiomas, esta nueva aplicación involucra activamente a sus usuarios para que contribuyan colaborativamente al diccionario universal de una manera fácil de usar y eficiente, con sugerencia de posibles traducciones basadas en transducciones de cambio de sonido de lenguajes relacionados y construcciones composicionales pan-lingüísticas.

P.5.28: Tagalog-English Code-Switching: Challenges for Automatic Detection

Authors: Nathaniel Oco

Country: Philippines

Abstract: In this poster, I will give a brief overview of the Philippines - a country in Southeast Asia - and discuss Tagalog-English code-switching (TECS). TECS is the use of both Tagalog and English in a discourse. I will also detail existing works to automatically detect TECS and the challenges ahead (e.g., intra-word code-switching and interlingual homographs).

Native language: Sa poster na ito, ako'y magbibigay ng maikling panimula sa Pilipinas - isang bansa sa Timog-silangang Asya - at aking tatalakayin ang Taglish. Ang Taglish ang paggamit sa parehas Tagalog at Ingles sa isang diskurso. Ako rin ay magbabahagi ng mga kaugnay na pananaliksik hingil sa awtomatikong pagsusuri sa Taglish at ang mga hamong kinahaharap (gaya ng panlalapi sa mga salitang Ingles at ang paggamit ng mga salitang parehas makikita sa Tagalog at Ingles).

P.5.29: How a low-resource named entities recognition and transliteration framework for Vietnamese can improve the automatic machine translation ?

Authors: Tan Ngoc Le and Fatiha Sadat

Country: Canada

Abstract: This presentation focuses on the low-resource pair of languages, French-Vietnamese, in order to develop a powerful machine translation system while focusing on the recognition of named entities and their transliterations. In addition to statistical approaches, we used a deep learning approach within our different systems to further improve the quality and efficiency of automatic translation of named entities and to reduce the rate of words outside vocabularies, untranslated and / or incorrectly translated words, but also to improve the quality of the machine translation system.

Native language: Bài trình bày này tập trung vào cặp ngôn ngữ tài nguyên thấp, tiếng Pháp-tiếng Việt, để phát triển một hệ thống dịch máy mạnh mẽ trong khi tập trung vào việc công nhận các thực thể được đặt tên và phiên âm của chúng. Ngoài các phương pháp thống kê, chúng tôi đã sử dụng phương pháp học sâu trong các hệ thống khác nhau của mình để cải thiện hơn nữa chất lượng và hiệu quả của dịch tự động các thực thể được đặt tên và để giảm tỷ lệ các từ bên ngoài từ vựng, từ chưa được dịch và / hoặc dịch sai, nhưng cũng để nâng cao chất lượng hệ thống dịch máy.

P.5.30: SITUATION AND CHALLENGES OF TECHNOLOGIES FOR INDIGENOUS LANGUAGES OF INDIA

Authors: Shweta Sinha and Shyam Sundar Agrawal

Country: India

Abstract: India is a country with huge linguistic diversity. Out of 900 languages spoken in the country, only a few have witnessed the digital world. This poster presents in detail the Indian languages situation in terms of resources,and technologies . It highlights the relative need , opportunities, barriers and complexities specific to the Indian Languages technologies. The aim is to study their influence on the adoption and adaptation of digital technology vis a vis Technological achievements/ fallout's of Indian languages relating to the world languages and to identify the gap and the need to take up future projects for technological advancements.

Native language: भारत एक विशाल भाषाई विविधता वाला देश है। देश में बोली जाने वाली 900 भाषाओं में से कुछ ही डिजिटल दुनिया में देखी गई हैं। यह पोस्टर, संसाधनों और प्रौद्योगिकियों के संदर्भ में भारतीय भाषाओं की स्थिति को विस्तार से प्रस्तुत करता है। यह भारतीय भाषाओं की प्रौद्योगिकियों के लिए विशिष्ट आवश्यकताओं, अवसरों, बाधाओं और जटिलताओं को उजागर करता है। इसका उद्देश्य डिजिटल प्रौद्योगिकी को अपनाने और विश्व भाषाओं से संबंधित भारतीय भाषाओं की एक तकनीकी उपलब्धियों को अपनाने, अंतर को पहचानने और तकनीकी प्रगति के लिए भविष्य की परियोजनाओं की आवश्यकता पर उनके प्रभाव का अध्ययन करना है।

P.5.31: Unicode for Indigenous Languages - Standards and technology for getting online
Authors: Craig Cornelius
Country: United States

Abstract: Three basic technologies are essential for a community to use its language with digital systems:

1. standardization defining digital codes for the writing system, 2.fonts and rendering technology for display and print
2. tools to create new text such as virtual keyboards and input methods

In addition, community leadership and engagement.

The poster discuss the Unicode Standard and associated technologies that are available today for use by indigenous communities. It will also outline steps that groups to make digital technology appropriate for their community needs.

P.5.32: CASS-LING's Linguistic Infrastructure: Resources, Platforms and Services
Authors: Wei Wang, Aijun Li and Danqing LIU
Country: China

Abstract: As China's highest academic institution of linguistic research, the Institute of Linguistics of Chinese Academy of Social Sciences (CASS-LING) has created language resources and platforms to provide varied services, which include: Contemporary Chinese Dictionary and Xinhua Dictionary with a Guinness World Record as the world's most popular dictionary; the Dictionaries and Speech Archives of Contemporary Chinese Dialects; an online system for dictionary compilation; the Database of the Grammars of the Chinese Dialects; a benchmark Putonghua pronunciation model; a visual 3D pronunciation model for English phonetic learning; the state government's examination and standardization of the pronunciation of characters and words.

Native language: 作为中华人民共和国语言学研究的最高学术机构，中国社会科学院语言研究所近年来搜集、整理大量语言资源，建设完成专业平台，提供以扎实学术研究为基础的社会服务。这些资源、平台、服务包括：《现代汉语词典》和《新华字典》（发行量吉尼斯世界纪录）；41卷本《现代汉语方言大词典》和40册《现代汉语方言音库》；词典编撰的支撑系统；已经搜集了10种方言数据的方言语法数据库；"九州音集"方言语音数据采集和分享微信平台；在调查4000多名儿童发音基础上推出的1.5~6岁普通话儿童语音发音常模；基于1万多小时不同方言区和少数民族地区英语学习者发音材料、面向英语语音教学的可视化3D发音模型；主导普通话审音工作。

P.5.33: Including Linguistic Knowledge in an Auxiliary Classifier CycleGAN for Corrective Feedback Generation in Korean Speech
Authors: Seung Hee Yang and Minhwa Chung
Country: Korea

Abstract: This work introduces a methodology to inject linguistic knowledge into an auxiliary classifier Cycle-consistent Generative Adversarial Network (CycleGAN), a machine learning method that retains domain knowledge. A related work used CycleGAN to generate a native version of a language learner's accented input speech in Korean. A linguistically-motivated auxiliary classifier is proposed in this work that enables generator-student interaction. This additional two-layer CNN learns to ensure the discriminability between the generated samples, and distinguishes three error types, "segmental," "suprasegmental," and "no correction" of the generated speech so that the learners will receive corrective feedback together with linguistic information.

Native language: 본 논문에서 제안하는 방법은 보조 분류기로 증강된 순환적 일관성 손실 함수를 사용한 생성적 적대 신경망 (Cycle-consistent Generative Adversarial Network; CycleGAN) 모델이며, 언어학 도메인 지식을 신경망 학습에 활용한다. 이전의 CycleGAN은 학습자의 목소리를 원어민 발음으로 변환하여 들려준다. 그러나, 이러한 피드백 방법은 어떠한 유형의 오류가 있었는지에 대한 정보를 제공하지 못한다는 데에 한계가 있다. 본 연구는 생성 결과 간의 언어학적 차별성을 학습한 보조 분류기를 기존의 CycleGAN에 추가하였다. 먼저 "분절음," "초분절음," "오류 없음" 유형별 CycleGAN 학습이 이루어지고, 모든 출력 결과는 하나의 분류기에 다시 입력된다. 최종적으로 교정 피드백은 교정 피드백 음성을 생성함과 동시에 해당 피드백이 어떤 오류인지 함께 제공한다.

P.6.1: Machine Translation 4 All: Developing informed and critical users through a program of machine translation literacy
Authors: Lynne Bowker
Country: Canada

Abstract: Machine translation is easy to use: copy, paste, click. However, just because users know HOW to use this tool doesn't mean that they are able to use it appropriately. In this age of free, online machine translation, we need for a new type of digital literacy: machine translation literacy. Literacy is a cognitive issue, rather than a techno-procedural issue. Machine translation literacy involves knowing not just how, but also whether, when and why to use machine translation. In Canada, we have developed guidelines that can help people who are not trained language professionals to become savvy users of machine translation.

Native language: La traduction automatique pour tous : Former des utilisateurs avertis et critiques avec une approche raisonnée de la traduction automatique

Les outils de traduction automatique sont faciles à utiliser : copier, coller, cliquer. Cependant, ce n'est pas parce que les utilisateurs savent COMMENT utiliser cet outil qu'ils sont bien équipés pour l'utiliser de façon optimale. À l'ère de la traduction automatique gratuite et en ligne, un nouveau type de culture numérique est nécessaire : il faut cultiver une « approche raisonner » de la traduction automatique. Développer une approche raisonnée, c'est une question cognitive plutôt qu'une question techno-procédurale. Une approche raisonnée de la traduction automatique implique de savoir non seulement comment, mais aussi si, quand et pourquoi utiliser la traduction automatique. Au Canada, nous avons élaboré des lignes directrices qui peuvent aider les personnes qui ne sont pas des professionnels langagières à devenir des utilisateurs avertis et critiques de la traduction automatique.

P.6.2: Building a common Digital Infrastructure to sustain Algonquian Languages
Authors: Marie-Odile Junker and Delasie Torkornoo
Country: Canada

Abstract: Our project includes Algonquian languages and communities of speakers, teachers and learners, at different degree of language vitality or endangerment (www.atlas-ling.ca, resources.atlas-ling.ca). Using a collaborative, participatory action research framework, we focus on dictionaries and integrated language resources. Our long-term goal is sustainability. Our design choices include:

- most affordable web server environments
- web frameworks that are reliable and diverse
- matured active open source frameworks
- systems capable of synchronizing online and offline data sources
- integration of the various applications We also raise questions about multimedia components, open-source, data stewardship, and long-term maintenance of not-for-profit resources.

Native language: Notre projet comprend des langues de la famille algonquienne présentant différents degrés de vitalité (www.atlas-ling.ca, resources.atlas-ling.ca). Dans un cadre de recherche particip-action, nous développons des dictionnaires et ressources linguistiques intégrées. Notre objectif est la durabilité. Nos choix de design incluent:

• environnements de serveur Web abordables

• cadres Web fiables et diversifiés

• infrastructures Open-Source matures et actives

• systèmes capables de synchroniser des sources de données en ligne et hors ligne

• intégration de différentes applications Nous soulevons des questions sur les composants multimédias, les logiciels Open-Source, la gestion responsable des données et la maintenance à long terme de ressources à but non lucratif.

P.6.3: On the promise and pitfalls of repurposing existing language technologies for endangered language documentation
Authors: Emily Prud'hommeaux, Robert Jimerson, Richard Hatcher, Raymond Ptucha and Karin Michelson
Country: United States

Abstract: Like many indigenous languages of North America, the Iroquoian language Seneca is endangered, with fewer than fifty living native speakers. Although descriptive grammars of Seneca exist, there are few texts and recordings available to support immersion programs and other revitalization efforts. In a collaboration between university researchers and Seneca community members, we are working to produce textual and audio documentation of the Seneca language using both existing toolkits and custom architectures. We find that while some toolkits yield promising results, the morphological complexity of Seneca and the variable quality of the available recordings present challenges for deploying one-size-fits-all solutions.

Native language: Dah ne:' dih tša'deyo'dëh oya' yei' niyöëdza'geh niodiwënö'dës koh neh onödowa'ga:' gawënö' agaiwahdö't so'jih gao' wis niwashëh niënödih ahsöh deodišnye'öh onödowa'ga:' gawënö'. Gwaheh ha'deyöh gayadö' niyo'dëh onödowá'ga:' gawënö' koh neh dohga:'ah niyoh gawënöhdas ogwenyöh aonödesdë' adiyë'he't onödowá'ga:' gawënö'. Dah ne:' hae'gwah dogëh dwadade'gë:' tënödeyësdahgwa'geh hënöjëönyanih koh neh onödowa'geonö' deodiyenö'. Dah ne:' hae'gwah hodihšöniaje' gayadöshäse:' watšowih na'ot gawënöhdas gayë' hadiyä'ta' dejaöh yesta' koh neh gadogëh neh nigayeëh gahšönih. Dah ne:' wa'agwaiwaho' neh dohga:'ah niyoh yesta' agwas wadesta' gwaheh so'jih ha'deyöh gayë' sgawënot koh neh dewenö:' niyo'dëh gawënohdas da'aöh ahsehšöni' gwisdë' neh ogwenyöh agaya'daei' gagwegöh koh neh agaiwaeis.

P.6.4: 7000 Languages: Free Language-Learning Software for Language Reclamation
Authors: Alexa Little, Kayleigh Jeannette and Kelsey Riggs
Country: United States

Abstract: 7000 Languages is a US non-profit organization that creates free language-learning software in partnership with Indigenous and endangered language communities. As of 2019, we have created 36 courses for 19 different languages, primarily Indigenous languages of North America. In this poster, we will show examples of our work and explain the process of creating a free online language course for your language.

Native language: 7000 Languages (7000 Lenguas) es una organización sin ánimo de lucra estadounidense que colabora con las comunidades indígenas y de lenguas en peligro de extinción para crear programas usadas para aprender esas lenguas. A partir del año 2019, hemos creado 36 cursos para 19 lenguas, que son principalmente de América del Norte. En este póster, vamos a mostrar ejemplos de nuestro trabajo y explicar el proceso de crear un curso libre en la red para la lengua suya.

P.6.5: National Research Council Canada Indigenous Language Technology Project
Authors: Aidan Pine, Nathan Brinklow, Heather Souter and Delaney Lothian
Country: Canada

Abstract: The Canadian federal budget released in 2017 invested $89.9M over three years to support Indigenous languages and cultures. The Canadian National Research Council (NRC) was granted $6M of this funding. Under the guidance of an Indigenous advisory committee, the NRC is researching and developing speech- and text-based technologies which aim to assist the stabilization, revitalization and reclamation of Indigenous languages.

We argue that successful technological solutions must support grass-roots language revitalization efforts and not overshadow them. This poster session describes three separate challenges faced by Indigenous language communities in Canada and highlights three corresponding NRC solutions.

Native language: Ne Canada kakó:ra óntka’we ne 2017 ohwísta, $89.9M nè:ne áhsen niiohserà:ke nika’shén:nes nè:ne aiakoié:nawa’se’ ne Onkwehón:we raotiwenna’shón:’a tánon tsi nihotirihó:ten. Ne Canadian National Research Council (NRC) $6M tahonwén:non e’tho tonterá:ko. Onkwehón:we rotitióhkwaien ne shakonaten’nikonhrá:wis ne enhonten’nikón:raren, ne NRC enhatirihó’kwate tánon enhonnón:ni ne tsi ní:ioht ne aiontá:ti tánon tsi ní:ioht tsi akahiatón:hake, á:se nahò:ten ronaterien’tatshén:rion tsi ní:ioht tsi akonwaié:nawa’se’ kwah tokén:en aiorihwahnírha, aonsontónhnhete tánon taontá:we ne Onkwehonwehnéha owenna’shón:’a. Iakwá:ton ne aioiá:neren’ne’ tsi nahatí:iere ne á:se ronaterien’tatshén:rion tsi nakaié:ren wentá:’on ki’ kwah tsi tka’nónhkwaien ne owén:na tsi sewatónhnhet, tóhsa ó:ia’ nahò:ten ohén:ton watohétstakw. Kí enkahiatonhserohá:rake áhsen niiorì:wake enwentó:ren’ne’ ne Onkwehonwehnéha owén:na aorihwà:ke ne tsi nón:we nikanakerahseraké:ron ne Canada, tánon ó:ni ronne’néstha ne áhsen niiorì:wake tsi nahatí:iere’ ne NRC.

P.6.6: St. Lawrence Island Language Technology for Documentation & Revitalization
Authors: Lane Schwartz, Emily Chen, Hayley Park, Sylvia Schreiner and Benjamin Hunt
Country: United States

Abstract: St. Lawrence Island Yupik is an endangered language indigenous to St. Lawrence Island in Alaska and the Chukotka Peninsula in Russia. This poster presents joint work in language documentation and computational linguistic research, designed to promote language revitalization efforts underway on St. Lawrence Island.

P.6.7: Building a Language Model of Nehiyawewin (Cree, Y-dialect)
Authors: Delaney Lothian, Daniela Teodorescu, Denilson Barbosa and Carrie Demmans Epp
Country: Canada

Abstract: Cree (Nehiyawewin) is a polysynthetic language Indigenous to the Canadian prairies. We are building a system intended to identifying morphemes using an N-gram language model of Cree. This research is motivated by focus groups that explored the needs and expectations of Indigenous people of Western Canada with respect to learning their ancestral language through technology. A lack of exposure to practice and limited resources are among the barriers they face. Our findings demonstrate the importance of including language structure as well as cultural meanings in technological tools to support learning ancestral languages instead of focusing only on vocabulary.

Native language: Le cri (Nehiyawewin) est une langue polysynthétique autochtone dans les Prairies canadiennes. Nous construisons un système destiné à identifier des morphèmes en utilisant un modèle de langage N-gramme du cri. Cette recherche est motivée par des groupes de discussion qui ont exploré les besoins et les attentes des peuples autochtones de l'Ouest canadien en ce qui concerne l'apprentissage de leur langue ancestrale par la technologie. Un manque d'exposition à la pratique et des ressources limitées font partie des obstacles auxquels ils sont confrontés. Nos résultats démontrent l'importance d'inclure la structure linguistique ainsi que les significations culturelles dans les outils technologiques afin de soutenir l'apprentissage des langues ancestrales au lieu de se concentrer uniquement sur le vocabulaire.

P.6.8: From Talking Leaves to Pixels: The Evolution of the Cherokee Syllabary
Authors: Roy Boney
Country: United States

Abstract: The Cherokee syllabary has been adapted to new technologies since its invention in 1821. It evolved from handwriting, to the printing press, to manual and electric typewriters, and now is part of the Unicode standard. This poster is a summary of the evolution of the Cherokee syllabary through the 21st century.

Native language: ᎾᏍᎩᏃ ᏣᎳᎩ ᏧᏃᏴᎩ ᎾᏳ 1821 ᏧᏕᏔᏴᏌᏗᏒ ᏗᎪᎷᏩᏛᏓ ᎾᏍᎩᎾᏃ ᏧᏩᎫᏔᏅᏒ ᎢᏤ ᏗᏕᏙᏈᏒ ᏗᏙᎳᎩ ᏂᏧᎵᏍᏔᏅᏍᏔᏅᎢ. ᎾᏍᎩᎾ ᎪᏪᎳᏅᎯ ᏅᏓᏁᏣᏴᏒ, ᎾᎢ ᏗᏕᎴᏴᏙᏗ ᏗᏐᎥᏍᏙᏗᎢ, ᎾᎢ ᏜᏩᏌ ᎠᏍᏕᏍᏗ ᎠᎴ ᎠᎾᏕᎵᏍᎩ ᏗᏐᏅᏍᏙᏗᏜ, ᎠᎴ ᏠᏤ ᏅᏌᏯᏍᏓ ᎾᏍᎩ ᎠᏰᎵ ᏕᏓᏂᏴᏛ ᏗᎪᏪᎵ ᎠᏟᏩᏍᏗ. ᎯᎠᏃ ᏕᏯᏑᏙᏗ ᏬᏃᏢᏍᎩ ᎾᏍᎩ ᏣᎳᎩ ᏧᏃᏴᎩ ᏕᏓᏁᏟᏴᏍᏔᏅ ᏧᎶᏒ ᎾᏍᎩ ᏐᏦᏁ ᏍᎪᎯᏧᏈ ᏧᏕᏘᏴᏓ.

P.6.9: ChoCo: A multimodal corpus for the Choctaw language

Authors: Jacqueline Brixey

Country: United States

Abstract: ChoCo is a general use corpus for Choctaw, an American indigenous language (ISO 639-2: cho, endonym: Chahta). The corpus contains audio, video, and text resources, with many texts also translated in English. The Oklahoma Choctaw and the Mississippi Choctaw variants of the language are represented in the corpus. The data set provides documentation support for the threatened language, and allows researchers and language teachers access to a diverse collection of resources.

Native language: ChoCo yvt chahta anumpa, miliki asha anumpa, ma i kanomma ish ia hinla. ChoCo yvt na hakll, na holbatoba apisa micha tali holisso anumpa ishi, awant anumpa lawa ho na hullo tosholi. ChoCo yvt Okla Homma micha Mississippi chahta im anumpa alhpesa holissochi chika ahobachi. ChoCo yvt anumpa mosholi ma holisso apela atahli micha na hoyo micha anumpa ikhanauchi ma tali holisso lawa atahli.

P.6.10: Issues and challenges of NLP in relation to Canada's Aboriginal languages

Authors: Fatiha Sadat, Tan Ngoc Le and David Huggins Daines

Country: Canada

Abstract: Natural Language Processing is a multidisciplinary field that aims to create tools and linguistic resources for various applications. These resources include emotion and sentiment analysis, speech analysis, machine translation, information extraction, prediction tools, and more. Through this presentation, we would like to present the issues and challenges of the NLP to endangered and / or poorly endowed languages such as Aboriginal languages. Also, we would like to present reflections on a multi-disciplinary project involving Aboriginal languages and cultures of Canada to build linguistic resources for machine translation and for learning and teaching Aboriginal languages and cultures.

Native language: هو مجال متعدد التخصصات يشمل اللغويات وعلوم الكمبيوتر والعلوم المعرفية. ويهدف إلى إنشاء الأدوات (NLP) لغة المعالجة الطبيعية والموارد اللغوية لمختلف التطبيقات. تتضمن هذه الموارد تحليل العاطفة والمشاعر ، وتحليل الكلام ، والترجمة الآلية ، واستخراج المعلومات ، وأدوات التنبؤ ، وأكثر من ذلك. من خلال هذا العرض التقديمي ، نود أن نعرض قضايا وتحديات البرمجة اللغوية العصبية على اللغات المهددة بالانقراض و / أو ذات اللغات الضعيفة مثل لغات السكان الأصليين. ونود أيضًا تقديم أفكار حول مشروع متعدد التخصصات يتضمن لغات وثقافات السكان الأصليين في كندا لبناء موارد لغوية للترجمة الآلية وللتعلم وتعليم لغات وثقافات السكان الأصليين.

Author Index

Organized by

in partnership with

ISBN: 979-10-95546-33-7
EAN: 9791095546337