Language technology for all: a challenge

Joseph J Mariani

4. Language technology for all: a challenge

Joseph MARIANI

Keywords: *language technology, spoken language processing, natural language processing, sign language processing, machine learning, artificial intelligence, human-machine interaction*

4.1 State of the Art

4.1.1 Key messages

Language technologies are mandatory to ensure language sustainability through multilingualism, aiming both at language preservation and communication. They are now widely spreading in everyday life through many applications, but they still need more research for more advanced applications, for better quality and especially for more languages.

Presently, between 100 and 150 languages benefit from language technologies, with very variable levels of quality. This represents only 2% of the more than 7,000 languages that are used in the world (please see Chapter 1.2 on language documentation for data, graphs and topics like language in cyberspace). **We should now find a way to address the remaining 98%, which include many indigenous languages**.

4.1.2 Definitions and descriptions

Language technologies (LT) cover a large area, starting from basic components such as a writing system, graphemic code, keyboard, phonetic alphabet, spelling checker or grammar checker.

Language technologies concern:

- Written language processing (also called natural language processing, or computational linguistics): text understanding and generation, summarization, search engines, information retrieval, answers to questions (Q&A), chatbots, text analytics, access to knowledge, sentiment and opinion analysis, and also machine translation and cross-lingual information retrieval in order to get access to information, whatever the language in which it has been encoded.
- Spoken language processing, which includes speech recognition and understanding, speech synthesis and generation, voice assistant, oral dialog, speaker recognition, emotion detection, and also language identification and speech interpretation.
- Sign language processing, in terms of analysis, generation and translation (Bragg et al. 2019).

Those cross-modal technologies are important in terms of accessibility, especially text-to-speech synthesis for people who are blind or visually impaired, speech-to-text transcription and sign language processing for deaf individuals, and voice command and control, including speech understanding and generation, for physically disabled people.

Speech interfaces are especially interesting as they only require simple, cheap, largely available cell phones and can be also used by the illiterate population or in the case of the many languages that do not have a writing system.

It is also important to stress that those technologies, such as machine translation or speech analysis and recognition, can be used as an aid to help people learning foreign languages, including pronunciation.

4.1.3 Facts and key findings

Quite critical is the fact that until we manage as humanity to sustain multilingual communication, we will not resolve the problem of language-based marginalization, as expressed in the sustainability and endangerment previous chapters. We need to move beyond the language-culture rationale to an argument based on language as a human right and a condition for human development.

The challenge of multilingualism is twofold: on the one hand to preserve cultures and their associated languages, and on the other hand to allow for communication across languages.

Most people want to preserve and keep using their language(s). Studies conducted in the European Union showed that 90% of the EU citizens prefer to have access to a website in their first language. But communicating is also a major challenge. The European Union has 24 official languages, e.g. 552 language pairs. The European Commission has a staff of more than 2,000 people taking care of translation across those languages, and the annual cost of multilingualism for the European Commission is estimated at more than 1 billion Euros. Despite this effort, multilingualism still appears as a major obstacle to the development of the European digital market and its 500 million customers, as it is estimated that one third of EU citizens would agree to buy goods on the internet in a foreign language, while 80% of EU citizens also think that websites in their language should be translated and made accessible to others who do not speak that language. It goes also with the international globalization of information: for example, more than 500 hours of new videos are uploaded every minute on You Tube, while more than 1 Billion hours of videos are watched everyday, in many different languages.

The needs related to multilingualism and crosslingualism exist in many different areas: for access to textual information in digital libraries, such as Europeana, with more than 60 million documents in more than 40 languages, or the UNESCO World Digital Library with more than 18,000 documents in more than 134 languages. For patents: the European Patent Office for example has a fund of close to 100 million patents in 32 languages, which would require an estimate of 300 million translations (that would need 1,500 years for a staff of 1,000 translators). Multilingualism is needed for technical notices, should it be for aeronautics, cars or domestic products, for automotive navigation, for interpretation in military or sanitary operations, often as a matter of urgency, as it was the case with the earthquake in Haiti, or for interpretation in the numerous conferences, meetings and lectures that are handled all over the world.

This clearly shows that it is impossible to answer quickly, or even to answer at all, to the numerous present and future needs for multilingualism with the present, and even future, human resources.

Considering multilingualism is not the first priority in any economic sector, but the sum of the small priorities in each of those economic sectors is large. It therefore necessitates a political thinking and a political action to merge all the small priorities into a large one.

Multilingualism is necessary, while its cost is very important and even exceeds available human resources. Given the multiple needs that are depicted and the large number of languages that are used worldwide, language technologies are therefore the only way to allow for generalizing multilingualism, if and only if the quality of language technologies meets the user needs (Mariani, 2014a)**.**

The effect is double:

- One may think that languages that lack language technologies will be less and less used: if people have to shift from their language to another language whenever they use a voice assistant or car navigation system, or play games, or make an emergency call, they will gradually stop using their language and shift to the language equipped with technology.
- On the contrary, languages that benefit from cross-lingual technologies, such as machine translation, will be more and more used, as they make it possible for someone to keep on using their language(s) while being understood by others. In everyday life, LT would make it possible to interpret in a doctor-patient exchange, to have identity documents issued in one's language, etc. We may also consider that if scientific papers would benefit from high quality translation, the situation would be different than the present one, where 96% of the *Web of Science* is in English, as researchers could keep writing articles in their language(s) while still being read and recognized by other researchers that do not use their language.

Language technologies have been incubating in research laboratories over the last half century (Mariani et al., 2019a and 2019b), and they are now widely spreading in everyday life through many applications, but they still need more research for more advanced applications, for better quality and especially for more languages (Mariani, 2019c).

Presently, 100 to 150 languages benefit from language technologies, less for the most advanced technologies (Fig. 4.1), with very variable levels of quality. This represents only 2% of the more than 7,000 languages that are used throughout the world. **The question is therefore: what shall we do for the remaining 98%?**

Voice Assistant	Languages	Language Variants
Amazon Alexa	English, French, German, Hindi, Italian, Japanese, Portuguese, Spanish	English (Australia, British, Canada, India, US), French (Canada, France), German, Hindi, Italian, Japanese, Portuguese (Brazil), Spanish (Mexico, Spain, US)
Apple Siri	Arabic, Chinese, Danish, Dutch, English, Finnish, French, German, Hebrew, Italian, Japanese, Korean, Malay, Norwegian, Portuguese, Russian, Spanish, Swedish, Thai, Turkish	Arabic, Chinese (Cantonese, Cantonese (Hong Kong), Mandarin, Mandarin (Taiwan)), Danish, Dutch (Belgium, Netherlands), English (Australia, British, Canada, India, Ireland, New Zealand, Singapour, South Africa, US), Finnish, French (Belgium, Canada, France, Switzerland), German (Austria, Germany, Switzerland), Hebrew, Italian (Italy, Switzerland), Japanese, Korean, Malay, Norwegian Bokmål, Portuguese (Brazil), Russian, Spanish (Chile, Mexico, Spain, US), Swedish, Thai, Turkish
Google Home	Arabic, Danish, Dutch, English, French, German, Hebrew, Hindi, Italian, Japanese, Korean, Norwegian, Portuguese, Spanish, Swedish	Arabic, Danish, Dutch, English (Australia, British, Canada, India, South Africa, US), French (Canada, France, Switzerland), German (Austria, Germany), Hebrew, Hindi, Italian, Japanese, Korean, Norwegian, Portuguese (Brazil), Spanish, Swedish
Samsung Bixby	Chinese, English, French, German, Italian, Korean, Portuguese, Spanish	Chinese (Mandarin), English (British, US), French (France), German (Germany), Italian (Italy), Korean, Portuguese (Brazil), Spanish (Spain)

Fig. 4.1 Example of languages covered by various Voice Assistants (as of June 11, 2020): from 8 to 20 languages and from 9 to 41 language variants.

The users of those languages are in an unbalanced situation. It creates a digital divide, and the corresponding languages are in danger of what can be called digital extinction, if not complete extinction.

4.1.4 Analysis and perspectives

In order to correct this digital divide and to provide equal rights to all, everyone should have the possibility to get access to workable language technologies in their language(s).

The needs are different, depending on the status of the language:

- Language documentation, for the languages that are not in use, or those that are close to being extinct due to the decreasing number of users (see Chapter 1.2 on documentation and description).
- Creation and development of basic language resources and technologies for the endangered languages (see Chapter 2.2 on endangerment).
- Development and deployment of LT-based applications, for the minority non-, or not-yet- endangered languages, in order to support their existence.
- Adaptation of already existing applications from the languages they exist in to other languages. This is the case for applications developed in the framework of projects such as human-machine communication, supported by the European Commission, which are often conducted for a small set of languages, while they should benefit all citizens of the European Union.
- Improvement of language technology quality, for most languages. This is obvious for machine translation, the quality of which is far from comparable to what is achieved by a human translator for most language pairs.
- Development of advanced language technologies, such as spoken dialog systems, human-robot interaction, or speech interpretation, for all languages including English.

We should therefore aim at providing language technologies for all: all languages, all people, all tasks, from language documentation to education and use in everyday life activities (health care, legal, information, administration, etc.) in which people are being marginalized due to language.

The question is then: How can we enlarge the presently limited linguistic coverage of language technologies?

Most of the present language technologies are based on machine learning, which is part of artificial intelligence. In order to achieve good quality, systems based on machine learning need huge amounts of language resources to be trained. The speech community use to say that: “there is no better data than more data”. The present speech recognition systems require around 10,000 hours of speech, from various speakers, of different gender, age and accent. Some speech corpora for English have now reached 100,000 hours. Similarly, the training of language models needs a billion words of texts, from various domains, including conversations. Some text corpuses for English have reached hundreds of billions of words. Machine translation is even more challenging as it requires parallel texts of human translations. Addressing translation for 7,500 languages means addressing more than 60 million language pairs and producing the corresponding parallel texts in sufficient quantity, while very few, or even no translation exists for most language pairs, and while there are many languages with no writing system.

Those data, also called language resources, include monolingual corpora (text, speech, images, videos), bilingual or multilingual parallel corpora for machine translation, lexica, electronic dictionaries, terminology databases etc. They are also necessary for research investigation in linguistics. Language resources are relatively easy to obtain for some languages, but much more difficult for others, the so-called **under-resourced or less-resourced languages** (see Chapter 1.2 on documentation and description).

In addition to data, the approach also needs system evaluation, in order to assess the quality of the systems, the progress made and whether the quality is adequate to the application needs. For some language technologies, metrics are well established, such as the *Word Error Rate* (WER) for speech recognition (Fig. 4.2a and 4.2 b).

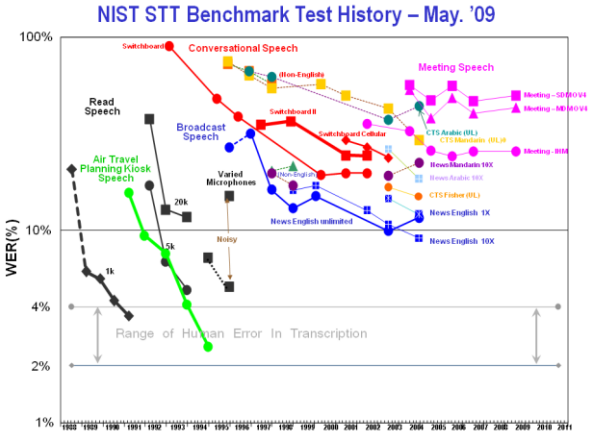
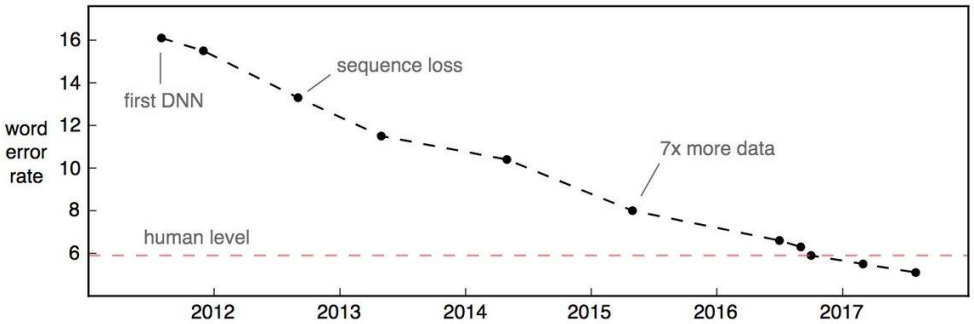


Figure 4.2a A history of Automatic Speech Recognition from 1987 to 2009 through the NIST evaluation campaigns[1]

This figure shows the progress of Automatic Speech Recognition over the years, through the international evaluation campaigns conducted by NIST. Shown on the chart are the best performances obtained that year, in terms of Word Error Rate (WER) in a logarithmic scale. The effort to go from 100% error (where the system does not recognise any word) to 10% is comparable to that required to go from 10% to 1% error rate. The tasks became increasingly difficult over the years (first with voice command, using an artificial language of 1,000 words, then voice dictation (20,000 words), radio/TV Broadcast News transcription (in English, Arabic and Mandarin Chinese) telephone conversations transcription (also in English, Arabic and Mandarin), meeting transcriptions), with variable conditions (real time or not, different qualities of sound recording). We see that for some tasks, the performance of systems is similar to those of a human listener, making these systems operational and marketable (such as for command languages). On the other hand, it is clear that for more complex tasks, performance improves more slowly, justifying the continuation of the research effort. Knowledge of these performances helps us to determine the feasibility of an application based on the quality level it requires. Thus, contrary to voice dialogue systems, an information retrieval system for audio-visual data does not require error-free performances in the transcription of speech, for example.



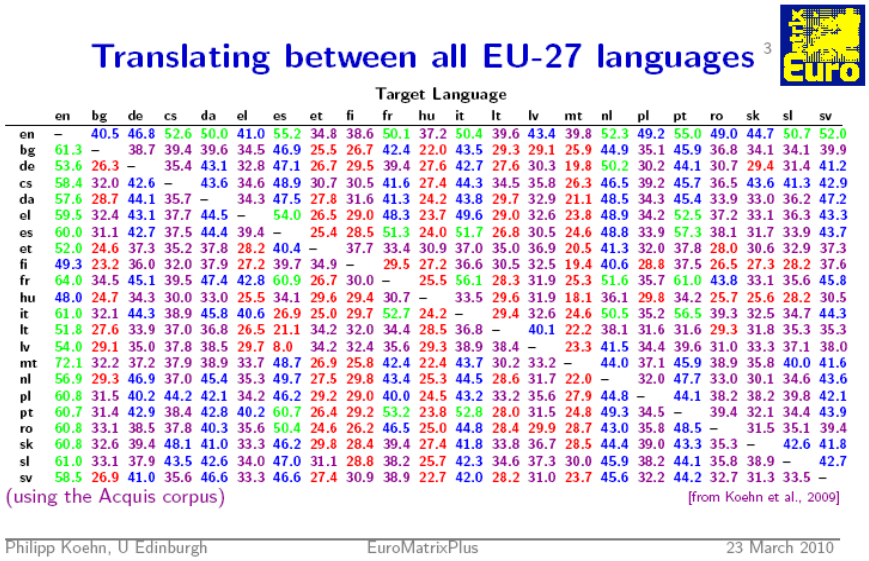
Improvements in word error rate over time on the Switchboard conversational speech recognition benchmark. The test set was collected in 2000. It consists of 40 phone conversations between two random native English speakers.

Fig. 4.2b Recent progresses on conversational speech recognition in English, showing that machine outpasses human in this specific task (Awni Hannun, 2017)[2]

[1] http://itl.nist.gov/iad/mig/publications/ASRhistory/index.html

[2] https://awni.github.io/speech-recognition/

For other technologies, they are not yet perfect, such as the BLEU (*bilingual evaluation understudy*) measure for machine translation (Papineni et al., 2000), which provides a percentage of similarities between the translation provided by the machine and a set of human translations of the same text (Fig. 4.3), or ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) for Text Summarization (Lin, 2004).



	en	bg	de	cs	da	el	es	et	fi	fr	hu	it	lt	lv	mt	nl	pl	pt	ro	sk	sl	sv
en	–	40.5	46.8	52.6	50.0	41.0	55.2	34.8	38.6	50.1	37.2	50.4	39.6	43.4	39.8	52.3	49.2	55.0	49.0	44.7	50.7	52.0
bg	61.3	–	38.7	39.4	39.6	34.5	46.9	25.5	26.7	42.4	22.0	43.5	29.3	29.1	25.9	44.9	35.1	45.9	36.8	34.1	34.1	39.9
de	53.6	26.3	–	35.4	43.1	32.8	47.1	26.7	29.5	39.4	27.6	42.7	27.6	30.3	19.8	50.2	30.2	44.1	30.7	29.4	31.4	41.2
cs	58.4	32.0	42.6	–	43.6	34.6	48.9	30.7	30.5	41.6	27.4	44.3	34.5	35.8	26.3	46.5	39.2	45.7	36.5	43.6	41.3	42.9
da	57.6	28.7	44.1	35.7	–	34.3	47.5	27.8	31.6	41.3	24.2	43.8	29.7	32.9	21.1	48.5	34.3	45.4	33.9	33.0	36.2	47.2
el	59.5	32.4	43.1	37.7	44.5	–	54.0	26.5	29.0	48.3	23.7	49.6	29.0	32.6	23.8	48.9	34.2	52.5	37.2	33.1	36.3	43.3
es	60.0	31.1	42.7	37.5	44.4	39.4	–	25.4	28.5	51.3	24.0	51.7	26.8	30.5	24.6	48.8	33.9	57.3	38.1	31.7	33.9	43.7
et	52.0	24.6	37.3	35.2	37.8	28.2	40.4	–	37.7	33.4	30.9	37.0	35.0	36.9	20.5	41.3	32.0	37.8	28.0	30.6	32.9	37.3
fi	49.3	23.2	36.0	32.0	37.9	27.2	39.7	34.9	–	29.5	27.2	36.6	30.5	32.5	19.4	40.6	28.8	37.5	26.5	27.3	28.2	37.6
fr	64.0	34.5	45.1	39.5	47.4	42.8	60.9	26.7	30.0	–	25.5	56.1	28.3	31.9	25.3	51.6	35.7	61.0	43.8	33.1	35.6	45.8
hu	48.0	24.7	34.3	30.0	33.0	25.5	34.1	29.6	29.4	30.7	–	33.5	29.6	31.9	18.1	36.1	29.8	34.2	25.7	25.6	28.2	30.5
it	61.0	32.1	44.3	38.9	45.8	40.6	26.9	25.0	29.7	52.7	24.2	–	29.4	32.6	24.6	50.5	35.2	56.5	39.3	32.5	34.7	44.3
lt	51.8	27.6	33.9	37.0	36.8	26.5	21.1	34.2	32.0	34.4	28.5	36.8	–	40.1	22.2	38.1	31.6	31.6	29.3	31.8	35.3	35.3
lv	54.0	29.1	35.0	37.8	38.5	29.7	8.0	34.2	32.4	35.6	29.3	38.9	38.4	–	23.3	41.5	34.4	39.6	31.0	33.3	37.1	38.0
mt	72.1	32.2	37.2	37.9	38.9	33.7	48.7	26.9	25.8	42.4	22.4	43.7	30.2	33.2	–	44.0	37.1	45.9	38.9	35.8	40.0	41.6
nl	56.9	29.3	46.9	37.0	45.4	35.3	49.7	27.5	29.8	43.4	25.3	44.5	28.6	31.7	22.0	–	32.0	47.7	33.0	30.1	34.6	43.6
pl	60.8	31.5	40.2	44.2	42.1	34.2	46.2	29.2	29.0	40.0	24.5	43.2	33.2	35.6	27.9	44.8	–	44.1	38.2	38.2	39.8	42.1
pt	60.7	31.4	42.9	38.4	42.8	40.2	60.7	26.4	29.2	53.2	23.8	52.8	28.0	31.5	24.8	49.3	34.5	–	39.4	32.1	34.4	43.9
ro	60.8	33.1	38.5	37.8	40.3	35.6	50.4	24.6	26.2	46.5	25.0	44.8	28.4	29.9	28.7	43.0	35.8	48.5	–	31.5	35.1	39.4
sk	60.8	32.6	39.4	48.1	41.0	33.3	46.2	29.8	28.4	39.4	27.4	41.8	33.8	36.7	28.5	44.4	39.0	43.3	35.3	–	42.6	41.8
sl	61.0	33.1	37.9	43.5	42.6	34.0	47.0	31.1	28.8	38.2	25.7	42.3	34.6	37.3	30.0	45.9	38.2	44.1	35.8	38.9	–	42.7
sv	58.5	26.9	41.0	35.6	46.6	33.3	46.6	27.4	30.9	38.9	22.7	42.0	28.2	31.0	23.7	45.6	32.2	44.2	32.7	31.3	33.5	–

Fig. 4.3 Machine Translation performances (BLEU measure) for 22 EU official languages, as of March 2010 (Koehn et al., 2009)

This figure gives the best performance obtained for 462 pairs of official languages of the European Union (lacking Irish Gaelic), in terms of their BLEU score (the higher the score, the better the translation, a human translator scoring around 80). The best results (shown in green and blue) correspond to the languages that benefit from research efforts in coordinated programmes, and from the availability of many parallel corpora (English, French, Dutch, Spanish, German,...), the worst (red) are languages that have not seen similar efforts, or that are very different from other languages (Hungarian, Maltese, Finnish ...).

For some, such as dialog systems, evaluation is still a research topic per se. Evaluation campaigns, also called benchmarking or shared tasks, consist in comparing the performances of different systems based on different approaches to common data, with the same protocol and agreed metrics. Training and test data, evaluation protocol, metrics and results can still be made available to all as "evaluation packages" after the evaluation campaign. The organization of international evaluation campaigns initiated by the US Department of Defense (DARPA) in the late 1980s (see Fig. 4.2) has been decisive in installing best practices in language processing research and allowing scientific progress, which has resulted in the availability of usable technologies. But also here, evaluation campaigns and packages only exist for a small set of languages, most notably English. Most results reflect the fact that more training data allows for better performances, as it appears in the case of Machine Translation for example (Fig. 4.4).

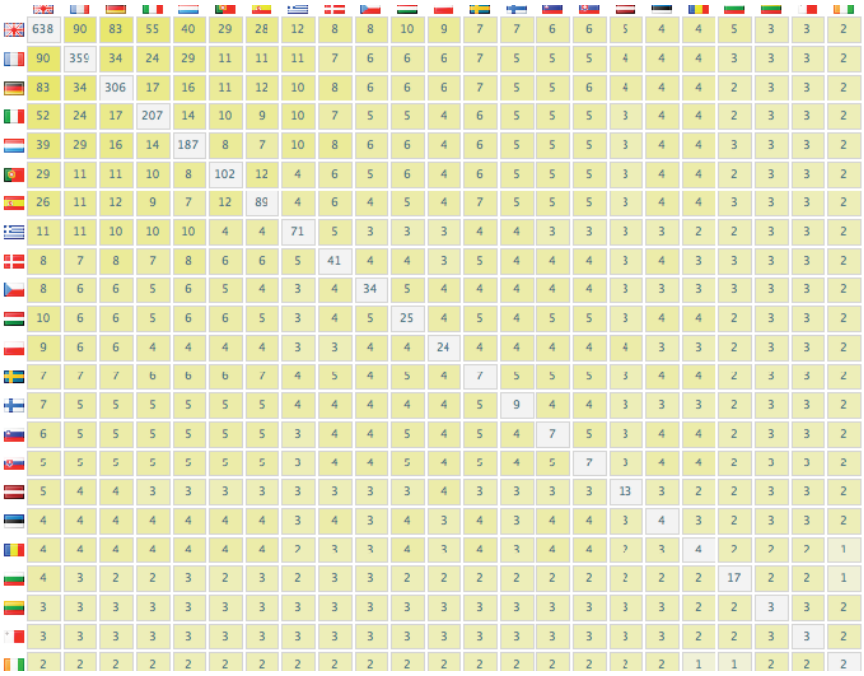


Fig. 4.4 Number of Parallel Corpora available for Machine Translation training related to the Machine Translation performances provided in Figure 4.3

The question is then: Is it possible to devote a similar effort for all languages, in terms of research investigations, language resources and technology evaluation, that was devoted for English, given that, for the past 50 years, the international research community has mostly worked and compared results for English, and much more modestly for a limited set of major international languages, such as French, Spanish or German.

The answer is that it is impossible for most languages, those that are referred to as "under-resourced languages", because of the cost it represents, and the lack of infrastructure, language resources and human resources. Moreover, translations don't exist for most language pairs and many languages (the so-called "oral languages"), as well as sign languages, do not have a writing system. For example, the transcription into phonemes of one minute of audio recordings requires one hour and a half of a linguist work (Austin and Sallabank, 2013).

4.2 Solution paths

Fortunately, it appears that there exist solution paths, as they were expressed by the participants of the LT4All conference[3,4], organized at the UNESCO Headquarters in Paris, between December 4-6, 2019 (see LT4All proceedings, forthcoming in 2020), **along various lines: organizational, economical, ethical, political and scientific**.

4.2.1 Organizational dimension

In order to achieve the goals of developing language technologies for more languages, there needs to be a shared effort that would include multiple stakeholders, from the public and private sectors. There are already:

- initiatives to provide international platform infrastructures where to find open source LT software and open data, such as the *European Language Grid* (ELG);
- several platforms where to gather and distribute data in many languages, such as:

[3] https://en.unesco.org/LT4All

[4] https://lt4all.lineupr.com/lt4all

  - *Common Voice* (Mozilla) for gathering data for speech recognition,
  - *Citizen linguist portal* (Linguistic Data Consortium , USA) for making citizens contribute as linguists,
  - *Language Sphere* (Ritsumeikan University , Japan) for the production of electronic dictionaries;
- and where to find tools, such as:
  - ELPIS (*Endangered Language Pipeline and Inference System*, from the Center of Excellence for the Dynamics of Language (CoEDL , Australia)),
  - *Deep Speech* (Mozilla) or *Kaldi* (Johns Hopkins University, USA) for speech recognition,
  - *Festival* (University of Edinburgh , UK) for text-to-speech synthesis,
  - *Cleo* (Amazon) for adapting the Alexa voice assistant to a new language,
  - *Laser* (*Language-Agnostic SEntence Representations*) from Facebook research,
  - *Systran NMT* (France), *Sockeye* (Amazon) or *Moses* (University of Edinburgh, UK, within the Euromatrix and TC-Star EC projects) for machine translation,
  - *Giellatekno* (Divvun, Arctic University of Norway) for processing morphologically-rich languages.

An *Open-LT manifesto* (openlt.org) is being circulated by Divvun asking that all language technologies be Open source: open localization, open interfaces, open resources and accessible standards.

In addition to data and tools, there is also a need for recipes, explaining how data and tools should be combined, for workshops and summer schools (such as JSALT (*Frederick Jelinek Memorial Summer Workshop on Speech and Language Technology*), Hackathons, Hand-on tutorials and training, especially through massive online open courses (MOOC), in order to facilitate mastering language technology for all researchers.

There should also be guides, on “How to get your language online?”, and there already exist descriptions of what should be a “Basic language resource kit” or a “Digital language survival kit”.

4.2.2 Economical dimension

The interest of IT companies was initially only for English: some of us remember the early time of computers where keyboards even did not include diacritics. It progressively extended to languages of rich countries, and some companies ranked countries according to their GDP and expressed their interest in developing language technologies for languages of countries ranked over a certain threshold. In a third period, the use of social networks extended the availability of language technologies for languages that have a lot of users. **The question is therefore now for languages of countries with low GDP and small populations**, that spark little interest for companies, given that 23 languages cover 50% of the world population[5], and the other half is covered by the remaining 6,977 languages or more. The answer may lie in the fact that some companies such as Facebook or Twitter are now committed to detect and block Fake News or harmful content on their social network. They have to do it quickly in order to avoid a large distribution, and they will therefore have to develop language technologies for the many languages in which they carry content.

The need for collaboration with the private sector is widely recognized, as the companies, and especially the GAFAM (Google, Apple, Facebook, Amazon, Microsoft) and the BATX (Baidu, Alibaba, Tencent, Xiaomi), have the necessary infrastructure to gather huge amounts

[5] https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers

of data through internet, smart home devices or car navigation applications that they already deployed.

A strong argument can be made for companies making an economic interest from wealthy people's data also having the moral duty to serve the needs of others.

4.2.3 Ethical dimension

As previously mentioned, the unavailability of language technologies will accelerate the extinction of the corresponding languages, given that some linguists think that 50% to 90% of languages may disappear by the end of this century if no action is taken (Austin and Sallabank, 2011).

Some indigenous peoples stress that language technologies should not be a weapon against their communities like school, church and media have been in the past, but should support their languages. They should be asked about their needs rather than being imposed the views of dominant parties on what those needs are, and their perspectives, which may be different from the mainstream views, should be taken into consideration. The property of data should remain with the language communities, even if it can be used by third parties.

In addition, ethics also includes other aspects such as privacy, trust, responsibility and the blurring between fiction and reality that must be avoided with the help of objective quality evaluation.

4.2.4 Political dimension

Initiatives have been taken to assess the situation regarding the availability of language technologies for the various languages used in a country or a group of countries. Such surveys have been achieved in the Language Technology Whitepapers produced by the T4ME project (*Technologies for a Multilingual Europe*) under the META-NET umbrella (*Multilingual Europe Technology Alliance Network*) for 31 European languages, including the 24 official ones. Information about the Language Technology level for each language is presented according to 4 LT related topics: Text processing, Speech processing, Machine Translation and Language Resources (corpus (text and speech), electronic dictionaries, etc.), based on the situation for the technologies and resources attached to those 4 topics according to various factors (such as Quantity, Availability, Quality, Coverage, Maturity, Sustainability, Adaptability) on a 5 to 0 scale (e.g. Excellent, Good, Moderate, Fragmentary, Weak, None) (Mariani et al., 2012) (Fig. 4.5).

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Language Technologies: Tools, Technologies, Applications							
Speech Recognition (voice command, voice dictation, broadcast transcription, conversational speech transcription, oral dialog)	4	3	4	4	4	3	3
Speech Synthesis (text-to-speech synthesis, speech generation)	4	3	4	4	4	3	3
Grammatical analysis (tokenization, POS tagging, morphological analysis/generation, shallow or deep syntactic analysis)	4	4	4	4	4	3	3
Semantic analysis (sentence semantics (WSD, argument structure, semantic roles, text semantics (coreference resolution, context, pragmatics, inference), Advanced Discourse Processing (text structure, coherence, rhetorical structure/RST, argumentative zoning, argumentation, text patterns, text types etc.)	3	3	3	3	3	2	2
Text Generation (sentence generation, report generation, text generation)	3	2	3	3	3	2	2
Machine Translation (and speech translation)	5	4	4	4	4	3	3
Language Resources: Resources, Data, Knowledge Base							
Text Corpora (Reference Corpora, Syntax corpora (treebanks, dependency banks))	4	3	4	4	4	4	3
Speech Corpora (raw speech data, labeled/annotated speech data, speech dialogue data)	4	3	4	4	4	4	3
Parallel Corpora, Translation Memory	4	3	4	4	4	4	3
Lexical Resources (Lexicons, Terminologies, Thesauri, WordNets)	4	3	4	4	4	4	3
Grammars (language models)	3	3	4	4	3	3	3

Figure 4.5 Estimated status of Language Technologies and Resources for the French language according to various factors (as of 2011) (Mariani et al., 2012)

This qualitative study was completed with a quantitative measure of the availability of language resources across languages (Mariani and Francopoulo, 2014b). The survey also comprised a comparison of language technology readiness according to natural language processing, speech processing, machine translation and language resources (Fig. 4.6). It concluded that 21 European languages were in danger of digital extinction (Rehm et al., 2016).

Domain	Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
Speech Processing		English	Czech, Dutch, Finnish, French, German, Italian, Portuguese, Spanish	Basque, Bulgarian, Catalan, Danish, Estonian, Galician, Greek, Hungarian, Irish, Norwegian, Polish, Serbian, Slovak, Slovene, Swedish	Albanian, Asturian, Bosnian, Breton, Croatian, Frisian, Friulian, Hebrew, Icelandic, Latvian, Limburgish, Lithuanian, Luxembourgish, Macedonian, Maltese, Occitan, Romanian, Romany, Scots, Turkish, Vlax Romani, Welsh, Yiddish
Machine Translation		English	French, Spanish	Catalan, Dutch, German, Hungarian, Italian, Polish, Romanian	Albanian, Asturian, Basque, Bosnian, Breton, Bulgarian, Croatian, Czech, Danish, Estonian, Finnish, Frisian, Friulian, Galician, Greek, Hebrew, Icelandic, Irish, Latvian, Limburgish, Lithuanian, Luxembourgish, Macedonian, Maltese, Norwegian, Occitan, Portuguese, Romany, Scots, Serbian, Slovak, Slovene, Swedish, Turkish, Vlax Romani, Welsh, Yiddish
Text Analysis		English	Dutch, French, German, Italian, Spanish	Basque, Bulgarian, Catalan, Czech, Danish, Finnish, Galician, Greek, Hebrew, Hungarian, Norwegian, Polish, Portuguese, Romanian, Slovak, Slovene, Swedish	Albanian, Asturian, Bosnian, Breton, Croatian, Estonian, Frisian, Friulian, Icelandic, Irish, Latvian, Limburgish, Lithuanian, Luxembourgish, Macedonian, Maltese, Occitan, Romany, Scots, Serbian, Turkish, Vlax Romani, Welsh, Yiddish
Language Resources		English	Czech, Dutch, French, German, Hungarian, Italian, Polish, Spanish, Swedish	Basque, Bulgarian, Catalan, Croatian, Danish, Estonian, Finnish, Galician, Greek, Hebrew, Norwegian, Portuguese, Romanian, Serbian, Slovak, Swedish	Albanian, Asturian, Bosnian, Breton, Frisian, Friulian, Icelandic, Irish, Latvian, Limburgish, Lithuanian, Luxembourgish, Macedonian, Maltese, Occitan, Romany, Scots, Turkish, Vlax Romani, Welsh, Yiddish

Fig. 4.6 Status of Language Resources and Language Technologies for 47 European languages[6], completed and updated in 2014 (Rehm et al., 2014))

The study was followed by a report committed by the European Parliament ("*Language equality in the digital age - Towards a Human Language Project*") (Rivera Pastor et al., 2017), which voted the resolution[7] "*Language equality in the digital age*" in September 2018 that resulted in a Call for Proposal "*Developing a strategic research, innovation and implementation agenda and a roadmap for achieving full digital language equality in Europe by 2030*" launched by the European Commission in June 2020[8].

A similar study is now conducted in India for the 22 scheduled languages (see Chapter on language diversity for a description of India's scheduled languages), and a survey was done in South Africa for the 11 official languages (see Chapter 3.1 on language emancipation for

[6] http://www.meta-net.eu/whitepapers/key-results-and-cross-language-comparison

[7] https://www.europarl.europa.eu/doceo/document/TA-8-2018-0332_EN.html

[8] https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/topic-details/langeq-2020;freeTextSearchKeyword=;typeCodes=1;statusCodes=31094501,31094502,31094503;programCode=PPPA;programDivisionCode=null;focusAreaCode=null;crossCuttingPriorityCode=null;callCode=PPPA-LANGEQ-2020;sortQuery=submissionStatus;orderBy=asc;onlyTenders=false;topicListKey=callTopicSearchTableState

details on the South African multilingual context) in the framework of the NHN (National Human Language Technology Network) (Grover et al. 2010), which resulted in the creation of the South African Centre for Digital Language Resources (SADiLaR). The European Federation of National Institutions for Languages (EFNIL) edits a European Language Monitor (ELM) every four years.

It is also worth noting that there exist national programs for language technologies, aiming at the internal national and regional languages in more and more countries, such as China, India (TDIL: *Technology Development for Indian Languages*), New Zealand, Iceland, Ireland, Wales, Norway (for the Sami language and also for Bokmål and Nynorsk), South Africa, Canada, USA, Mexico, Morocco (for the Tamazight language at IRCAM), or at a set of languages spoken over a continent (European Union, Africa (Masakhane project)), but also programs in the US, which were established for external affairs be it for geopolitical or humanitarian reasons.

Nonetheless, it is reported that domestic funding is not sufficient, and that an international coordination through UNESCO would be needed[9], in partnership with other stakeholders from the public and private sectors.

4.2.5 Scientific dimension

There are also good news coming from science, as more and more research activities are devoted to under-resourced languages.

On the mainstream traditional **Machine Learning** approach, efforts have been devoted at enlarging the size of language datasets in many countries that decided to support their language(s), as it appears in a Language Resources survey such as the LRE Map (Calzolari et al., 2010) or at organizing evaluation campaigns, such as ELLORA (*Enabling Low Resource Languages*) for the Indian languages (Mohan et al., 2018).

As an alternative to **machine learning**, several teams work on **rule-based approaches** for morphologically rich languages that lack language resources, such as *Divvun* for the Sami languages.

Bootstrapping approaches are tried out for the development of dialog systems, which consist in providing a dialog system with low, yet sufficient quality that is improved over time through the use of the system, and which allows for enlarging the training corpus and hence improving the performances.

It is also possible to work on **several languages of the same family**, including the dialectal varieties that are mostly expressed by variants in the lexicon or in the accent.

The case of **oral languages** without any writing system can be approached by recording native speakers in their usual environment, **re-speaking** their talks in good acoustic conditions with a single speaker who also translates them in a language benefiting from language technologies (Bird, 2010). It is then possible to train a speech recognition system through machine learning on that data. The system can then be used to also study the language, while the translation in a well-resourced language can enlarge the size of the research community addressing this oral language (Adda et al., 2018).

Transfer learning allows for porting the models developed for one language with sufficient training data to another (similar) one that may have less data.

In the **multilingual approach**, several languages are considered together. In speech recognition, for example, the acoustic models of the phonemes of several languages may be used to

[9] A statement by Satoshi Nakamura at the LT4All conference.

recognize a new language. In machine translation, instead of training a translation engine on the parallel corpus of two language pairs, it is possible to train it on all available language pairs (Schwenk and Douze, 2017). Interestingly, some results show that multilingual language models are better than monolingual language models on monolingual applications, or than bilingual models in machine translation.

In order to increase the size of the training data, it is also possible to artificially produce **augmented data** through language generation or speech synthesis. For example, speech at different speeds or in various noisy conditions can be artificially produced from an initial sample.

In the **speech chain** approach (Tjandra et al., 2017), bi-directional speech-to-text (STT) and text-to-speech (TTS) modules are used in a loop to gradually improve the quality of both modules: the text resulting from the transcription of a speech signal is synthesized and the signal resulting from this synthesis is compared with the original. The parameters are modified in order to minimize this difference. Methods of reinforcement learning and adversarial learning can be used in this framework.

In machine translation, the **zero-shot approach** (Johnson et al. 2017) experiments showed that it is possible to achieve Japanese-Korean translations without any Japanese-Korean parallel texts, by merging previously developed English-Japanese and English-Korean translation systems. The neural networks corresponding to those two translation systems are cut in two parts and pasted, which may induce that the pasted cells represent language-independent meaning.

The **zero-resource approach** (Jansen et al. 2012) aims at mimicking the way children learn to speak without any lexicon, grammar and writing system, but using multimodal communication, including visual and gesture information. Experiments have been conducted on speech recognition and speech synthesis with this approach.

The ultimate approach may reside in **language understanding**, as languages use different sounds, different words, different syntaxes to express the same meaning. Portability across languages would be easy if it would be possible to extract meaning from the linguistic content in one language, and to generate linguistic content from that meaning in another language.

4.2.6 Challenges

Several scientific challenges will have to be addressed in the coming decade and constitute open research problems at various extents.

The processing of **prosody** (accent, intonation and rhythm) is still relatively unresolved, as it is difficult to annotate this so-called “supra-segmental” information in the signal, while it carries nuances that express emotions, for example. **Tone languages** also require the study of supra-segmental information. **Dialectal variations** must be studied, and they are numerous in some languages, such as Arabic. Systems must deal with **code-switching** and **code-mixing**, when two or more different languages are used in the same sentence. The processing of **oral languages** is especially difficult, as there is no transcription into written words of the speech signal that contains no pause between words and no punctuation marks at the end of sentences, but also quite appealing as it would allow access to digital content, including oral tradition, through keyword spotting. **Spoken or written** dialog is still an open problem as it means processing information that is shared by two interlocutors or more in the discussion, sometimes expressed as indirect speech acts. As dialogs are dynamic processes,

it is difficult to define methodologies and metrics that can assess and compare system quality.

Taking into account the **contextual**, non-verbal information is another challenge, as well as **common-sense** information that can hardly be learned from textual training data, and consciousness specific to humans, but not to virtual agents. The processing of **sign languages** adds another challenge, as it necessitates to also address visual scene analysis and generation, not only of the hands but of the full body, which is a research area per se. **Fake news and harmful content detection** constitutes a new reason for companies marketing social networks to address a larger set of languages. If they agree to be committed to avoid distributing unlawful content, they need to be able to detect such content as soon as it is posted, and the only way to do so is automatic information extraction in real time in many languages. **Human-robot interaction** raises many problems related to various communication modalities through sound, vision and gestures while one or several humans and robots move in an open environment. **Automatic interpretation** is not simply speech recognition in the source language, followed by machine translation, followed by speech synthesis in the target language, but requires language understanding, summarization capacity and language generation in order to operate in real time with sufficient quality, both in terms of adequacy and fluency.

If we may consider in a first analysis that language understanding is a key issue for allowing the development of language technologies in many languages, it appears that the meaning of a phrase may differ according to the various cultures corresponding to the various languages **(diversity in meaning)**. It is well known that languages spoken in the arctic regions have up to 50 different words for expressing what is just expressed by “snow” in the temperate regions, and that some words in a language are considered as non-translatable in another languages. This problem should be handled through paraphrases or explanations.

4.2.7 Conclusion and Recommendations

In conclusion, we believe that the availability of language technology is crucial for the sustainability of a language and that the situation is very different between the major languages, and especially English, used by a large population, and the long tail of community languages, including indigenous languages.

Science may bring solutions to this problem and research should be supported at a higher pace by multi-stakeholders, including national and regional public institutions in the various countries, international bodies such as the European Union, and the private sector. UNESCO could play a major role in coordinating this effort at the international level.

Conclusions, recommendations and suggested actions related to language technologies appear in Conclusion V of the Strategic Outcome Document of the 2019 International Year of Indigenous Languages[10], that apply not only to indigenous languages but to all marginalized languages, i.e. probably all languages apart from English:

Conclusion V.

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 from older to younger generations, rather than fostering their disappearance 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 that impede the availa-

[10] https://unesdoc.unesco.org/ark:/48223/pf0000371494?posInSet=1&queryId=31e33472-617a-4e80-a0cd-5f144161f06f - Paris, November 2019

bility 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 including in particular emerging technologies, Artificial Intelligence, Blockchain and others ***(PROGRESS)****.*

Goal (V)

Indigenous peoples should have the possibility to benefit from the full range of language technologies which help people to break through the potential barriers of the digital divide, giving them open access to, and production capability in, multilingual knowledge and educational materials, along with the benefit of available public services in their own languages.

Recommendations:

5.1. Member States, should draw on the best information, know-how and methodology at their disposal when ***formulating and planning how they will implement and evaluate inclusive language policies*** *and should take effective measures to ensure that scientific and technological developments are leveraged for the benefit of users of each language, and that they address the situation of individual languages and their users, guaranteeing equal rights to be educated, to inherit their traditional culture, and to enjoy the service benefits and convenience of modern technological products, in whose design, development and production they should, so far as possible, be engaged.*

5.2. Member States and other relevant stakeholders, in full collaboration with indigenous peoples, should ***use Information and Communication Technologies (ICT) including Artificial Intelligence (AI), Blockchain and other, to promote the creative transformation, innovative development and effective dissemination of language resources and seek new ways to protect and respect the indigenous traditional knowledge****.*

5.3. Building where possible on existing work, ***a set of international standards should be developed and agreed to protect essential language resources, in cooperation with indigenous language users****; these must cover (i) technical standards for collection, annotation and documentation and (ii) collaboration procedures in the construction, sharing and application of language resource standards globally. International standards organizations and professional bodies (universities, research institutes, individual experts and other stakeholders) have the responsibility to be engaged in language protection and preservation, to in the first instance formulate and thereafter to uphold the agreed standards.*

5.4. Member states and other stakeholders, in close collaboration with indigenous peoples, should ***develop advanced tools for the collection and analysis of language data as well as for the transliteration and annotation of multi-modal content collections and cultural exhibitions where such do not already exist.*** *This will allow the development of technologies specifically adapted to the characteristics of indigenous languages, which in turn will strengthen and underpin the status of these languages; such tools include speech recognition, synthesis systems and machine translation technology.*

Suggested actions (V):

a) ***Develop teaching facilities, techniques and devices specifically for the purpose of supporting indigenous languages****, of designing educational curricula and of supplying necessary tools for advanced translation, using Artificial Intelligence and machine learning techniques,*

b) ***Integrate and share successful language acquisition and learning techniques as well as intergenerational transmission methodologies****, including language immersion and bilingual education methods, thereby supporting quality learning environments, the principle of equal and inclusive access for all, and the training of new teachers and of those already in service,*

c) ***Provide access to funding sources for the research projects of indigenous people****s, seeking to reconcile potentially the competing priorities of academia and indigenous peoples,*

d) ***Establish and support institutional structures for indigenous languages monitoring, evaluation and impact****, led by, and developed in collaboration with, indigenous peoples,*

e) ***Encourage collaboration between industry, the research and development sector and indigenous peoples****, focusing on the needs and interests of indigenous communities, extending and refining current language technologies as well as designing new ones, and developing necessary algorithms, applications and systems to support indigenous peoples in their own use of the internet and social media networks. The Artificial Intelligence paradigm should be rolled out within an ethical framework.*

Conclusion V of the Strategic Outcome Document
of the 2019 International Year of Indigenous Languages

In this perspective, it is proposed to install a language technology committee, which would be attached to the Decade of Indigenous Languages 2022-2032 Steering Committee. In order to assess the progress made in this regard, it is proposed that regular conferences similar to LT4All be organized, gathering representatives from the technological and political spheres connected to languages. In order to assess progress, it is proposed to regularly measure the increase in the number of languages that benefit from workable language technologies by monitoring core indicators, similar to what has been done in Europe, India and South Africa. In order to get this information, a survey similar to the META-NET Whitepapers for Europe could be extended to the international scene, or it could be harvested in the framework of the questionnaire of the *World Atlas on Languages* (WAL) under development at UNESCO.

References

- Adda, Gilles / Adda-Decker, Martine / Ambouroue, Odette / Besacier, Laurent / Blachon, David / Maynard, Hélène / Godard, Pierre / Hamlaoui, Fatima / Idiatov, Dmitry / Kouarata, Guy-Noël / Lamel, Lori / Makasso, Emmanuel-Moselly / Mariani, Joseph-Jean / Rialland, Annie / Stucker, Sebastian / Van de Velde, Mark / Gauthier, Elodie / Yvon, François / Zerbian, Sabine (2018). BULB: Breaking the Unwritten Language Barrier), In Computational Methods for Endangered Language Documentation and Description, 2018.

- Austin, P.K. / Sallabank, J. (2011). "Introduction". In Austin, Peter K; Sallabank, Julia (eds.). Cambridge Handbook of Endangered Languages. Cambridge University Press

- Austin, P.K. / Sallabank, J. (2013). Endangered languages: an introduction. Journal of Multilingual and Multicultural Development. Volume 34, 2013 - Issue 4: Endangered Languages, Taylor & Francis.

- Bird, Steven (2010). A scalable method for preserving oral literature from small languages. In Proceedings of the 12th International Conference on Asia-Pacific Digital Libraries.

- Bragg, Danielle / Koller, Oscar / Bellard, Mary / Berke, Larwan / Boudrealt, Patrick / Braffort, Annelies / Caselli, Naomi / Huenerfauth, Matt / Kacorri, Hernisa / Verhoef, Tessa / Vogler, Christian / Morris, Meredith (2019). Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective.

https://www.researchgate.net/publication/335395194_Sign_Language_Recognition_Generation_and_Translation_An_Interdisciplinary_Perspective

- Calzolari, Nicoletta / Soria, Claudia / Del Gratta, Riccardo / Goggi, Sara / Quochi, Valeria / Russo, Irene / Choukri, Khalid / Mariani, Joseph / Piperidis, Stelios (2010). The LREC Map of Language Resources and Technologies. LREC-2010, Malta

- Grover, Aditi Sharma / van Huyssteen, Gerhard B. / Pretorius, Marthinus W. (2010). The South African Human Language Technologies Audit, Language Resources and Evaluation Conference, Malta, May 17-23 2010

- Jansen, A. / Dupoux, E. / Goldwater, S. / Johnson, M. / Khudanpur, S. / Church, K. / Feldman, N. / Hermansky, H. / Metze, F. / Rose, R. / Seltzer, M. / Clark, P. / McGraw, I. / Varadarajan, B. / Bennett, E. / Boerschinger, B. / Chiu, J. / Dunbar, E. / Fourtassi, A. / Harwath, D. / Lee, C.y. / Levin, K. / Norouzian, A. / Peddinti, V. / Richardson, R. / Schatz, T. / Thomas, S. (2013). A summary of the 2012 JH CLSP Workshop on zero resource speech technologies and models of

early language acquisition. In *ICASSP-2013 (IEEE International Conference on Acoustics Speech and Signal Processing)*, (pp 8111-8115)

- Johnson, Melvin / Schuster, Mike / Le, Quoc V. / Krikun, Maxim / Wu, Yonghui / Chen, Zhifeng / Thorat, Nikhil / Viégas, Fernanda / Wattenberg, Martin / Corrado, Greg / Macduff, Hughes / Dean, Jeffrey (2017). Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation, Transactions of the Association for Computational Linguistics, vol. 5, pp. 339–351, 2017

- Koehn, Ph. / Birch A. / Steinberger R. (2009). 462 Machine Translation Systems for Europe, Machine Translation Summit XII, p. 65-72, 2009.

- Lin, Chin-Yew (2004). ROUGE: a Package for Automatic Evaluation of Summaries. In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), Barcelona, Spain, July 25 - 26, 2004

- Mariani, J. / Paroubek, P. / Francopoulo, G. / Max, A. / Yvon, F. / Zweigenbaum,P. (2012). The French Language in the Digital Age / La Langue Française à l'Ere du Numérique, 102 pp., Springer, 2012, ISSN 2194-1416, ISBN 978-3-642-30760-7

- Mariani, J. (2014a). How Language Technologies Can Facilitate Multilingualism, in Linguistic and Cultural Diversity in Cyberspace, Proceedings of the 3nd International Conference (Yakutsk, Russian Federation, 30 June – 3 July 2014), July 2015, pp 48-60, ISBN 978-5-91515-063-0

- Mariani, J. / Francopoulo, G. (2014b). Language Matrices & the Language Resource Impact Factor, in Language Production, Cognition, and the Lexicon, Festschrift in honour of Michael Zock, N. Gala, R. Rapp, G. Bel eds., Springer, ISBN 978-3-319-08042-0, December 2014

- Mariani, Joseph / Francopoulo, Gil / Paroubek, Patrick (2019a). The NLP4NLP Corpus (I): 50 Years of Publication, Collaboration and Citation in Speech and Language Processing., *Frontiers in Research Metrics and Analytics,* "Mining Scientific Papers: NLP-enhanced Bibliometrics" Iana Atanassova, Marc Bertin and Philipp Mayr eds, Vol. 3, 36 pages, 2019, DOI= 10.3389/frma.2018.00036, ISSN=2504-0537
https://www.frontiersin.org/article/10.3389/frma.2018.00036

- Mariani, Joseph / Francopoulo, Gil / Paroubek, Patrick / Vernier, Frédéric (2019b). The NLP4NLP Corpus (II): 50 Years of Research in Speech and Language Processing, *Frontiers in Research Metrics and Analytics,* "Mining Scientific Papers: NLP-enhanced Bibliometrics" Iana Atanassova, Marc Bertin and Philipp Mayr eds, Vol. 3, 37 pages, 2019, DOI= 10.3389/frma.2018.00037, ISSN=2504-0537,
https://www.frontiersin.org/article/10.3389/frma.2018.00037

- Mariani, Joseph (2019c). Language Technologies in Support to Multilingualism, World Summit on Information Society Forum, Geneva, April 10, 2019

- Mohan, Brij / Srivastava, Lal / Sitaram, Sunayana / Bali, Kalika / Mehta, Rupesh Kumar / Mohan, Krishna Doss / Matani, Pallavi / Satpal, Sandeepkumar / Srikanth, Radhakrishnan / Nayak, Niranjan (2018). Interspeech 2018 Low Resource Automatic Speech Recognition Challenge for Indian Languages, Spoken Language Technology for Under-resourced Languages conference (SLTU 2018), Gurgaon, August 2018

- Papineni, Kishore / Roukos, Salim / Ward, Todd / Zhu, Wei-Jing (2002). BLEU: a Method for Automatic Evaluation of Machine Translation, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), 2002, pp. 311-318

- Rehm, Georg / Uszkoreit, Hans / Ananiadou, Sophia / Bel, Núria / Bielevičienė, Audronė / Borin, Lars / Branco, António / Budin, Gerhard / Calzolari, Nicoletta / Daelemans, Walter / Garabík, Radovan / García-Mateo, Carmen / van Genabith, Josef / Hajič, Jan / Hernáez, Inma /

Judge, John / Koeva, Svetla / Krek, Simon / Krstev, Cvetana / Lindén, Krister / Magnini, Bernardo / Mariani, Joseph / McNaught, John / Melero, Maite / Monachini, Monica / Moreno, Asunción / Odjik, Jan / Ogrodniczuk, Maciej / Pęzik, Piotr / Piperidis, Stelios / Przepiórkowski, Adam / Rögnvaldsson, Eiríkur / Rosner, Mike / Sandford Pedersen, Bolette / Skadiņa, Inguna / De Smedt, Koenraad / Tadić, Marko / Thompson, Paul / Tufiş, Dan / Váradi, Tamás / Vasiļjevs, Andrejs / Vider, Kadri / Zabarskaite, Jolanta (2016). The Strategic Impact of META-NET on the Regional, National and International Level, Language Resources and Evaluation Journal, 2016, pp 1-24, ISSN: 1574-0218, doi: 10.1007/s10579-015-9333-4

- Rehm, Georg / Uszkoreit, Hans / Dagan, Ido / Goetcherian, Vartkes / Dogan, Mehmet Ugur / Mermer, Coskun / Varadi, Tamás / Kirchmeier-Andersen, Sabine / Stickel, Gerhard / Prys Jones, Meirion / Oeter, Stefan / Gramstad, Sigve (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, May 2014.

- Rivera Pastor, Rafael / Tarín Quirós, Carlota / Villar García, Juan Pablo / Badia Cardús, Toni / Melero Nogués, Maite (2017). Language equality in the digital age - Towards a Human Language Project, Study IP/G/STOA/FWC/2013-001/Lot4/C2 (March 2017).

- Schwenk, Holger / Douze, Matthijs (2017). Learning Joint Multilingual Sentence Representations with Neural Machine Translation, 2nd ACL Workshop on Representation Learning for NLP, Vancouver, August 3rd 2017.

- Tjandra, Andros / Sakti, Sakriani / *Nakamura,* Satoshi *(2017)*. Listening while Speaking: Speech Chain by Deep Learning, ASRU 2017, Okinawa, December 16-20, 2017