Preface

Joseph J Mariani, Zygmunt Vetulani

Preface

We began the series of the LTC conferences at a time when Europe was preparing for the "Great Enlargement" of Central European countries, which in particular involved Poland. At that time not only for visionaries, but also for leaders and enlightened European politicians, the terms "globalization" and "Information Society" were gaining in popularity. Globalization as a leading trend of technological development, Information Society as an environment and at the same time the recipient of global products.

The concept of an information society assumes general circulation of information beyond language and cultural barriers while preserving multicultural heritage as a guarantor of the preservation of European multilingualism not only in Europe but on a greater global scale. When in 1995 the co-author of this preface, Zygmunt Vetulani, organized the LTC for the first time, it was difficult to predict that a two-day meeting of Polish linguists and computer scientists with experts from the European Commission was a kick-off event of an international conference series to be organized systematically over more than 20 years.

The LTC 1995 meeting was organized in the context of the "Language and Technology: Awareness Days", a series of European Commission events (DG XIII) located in various European countries, and the abovementioned assumptions recurred in the presentations of eminent experts including Jan Roukens, Antonio Zampolli and Dafydd Gibbon. The programmatic paper by Jan Roukens emphasized the concern of controlling globalization in a civilized manner. Roukens recalled the findings of the summit in Corfu (June 1994) regarding multilingualism. *"The people of the Union, with their different languages, cultures, history and educational systems, should be able to communicate with each other and the external world in ways that allow them to live and work together in an effective, productive, tolerant, democratic and cohesive way in their common European 'home'."* As a dominant technology, the European Commission advocated – and still advocates – machine translation (Stroerup and Maegaard) as a development paradigm, which a visionary of the field, Antonio Zampolli, postulated for the development of a human language technologies industry. This industry has developed far beyond the scope of these technologies in 1995. At that time, Microsoft was already 20 years old, Apple Computer Inc. was 19, the IBM PC was just 14 years old, the Apple Macintosh was 12 years old, the World Wide Web was only 6 years old, and there was no Google, no Amazon; portable computers still weighted as much as sawing machines and there were no internet-capable devices like tablets or smartphones. In contrast, the current state of the art includes tablets and smartphones, these small, lightweight ubiquitous universal multimedia devices for gaming, searching, messaging, socially interacting, video conferencing as well as phoning.

The enlargement of the European Union on 1st May 2004 created a further incentive to draw lessons from the success of the LTC in 1995, which was attended by almost all active researchers in the field of language technology in Poland, and to convene a regular conference on a European scale. We have undertaken the organization of the LTC conference in Poznań every two years and we consistently fulfil the vision of 1995, as evidenced by its 20th anniversary in 2015. In the period from 1995 until now, an extensive international community has been created, which, at the time we write (2018), exceeds 1200 authors from more than 60 countries around the world. Similarly, the global reach has a community of program committee members and reviewers who uphold the high scientific level of the conference presentations. At the same time, we put the emphasis on promoting the best work of young researchers by establishing awards for the best student papers, as well as trying to encourage the continuation of studies recognized as outstanding in the qualification process by creating, in a productive cooperation with the Springer publishing house, a series of monographs

published in the LNAI series where selected conference works are published as *Revised Extended Papers*, including continuations of the research results presented at the LTC.

The LTC 2015 conference was dedicated to the memory of Adam Kilgarriff, who passed away in 2015. Adam Kilgariff brought major contributions in the fields of lexical semantics and corpora. He especially allowed for achieving scientific advances in the semantic disambiguation of word meaning through the organization of evaluation campaigns.

The analytical development of the LTC achievements until 2015 was depicted at the conference in the form of a plenary presentation entitled "*Rediscovering 20 Years of Discoveries in Language & Technnology"* (Mariani, Paroubek, Francopoulo, and Vetulani), which navigates through the 555 papers and the corresponding 959 authors who contributed to the conference since its launching in 1995 (paper accessible from http://ltc.amu.edu.pl/a2015/).

As a continuation of the series initiated in 2009, LTC 2015 expressed the same interest for Less-Resourced languages within a special session devoted to that topic, which addressed languages such as Malagasy, Vietnamese, Sambalpuri, Swiss German, Irish, Scottish Gaelic or Welsh, but also Sanskrit or Ancient Greek. An Invited Talk presented by Dafydd Gibbon, specifically stressed the case of Endangered Languages, while Mikel L. Forcada described the situation regarding Machine Translation of such languages.

Regardless of the impressive achievements of the last 20 years in many areas, including the development of multimedia technologies, the integration of speech and language technologies, and information and communication technologies for an increasing number of languages, we must notice that building a global information society does not bring only advantages, but also dangers. The need for civilized control of the complex processes of creating a global information society has been demonstrated by many violent attacks on civilization as we know, endangered by the activities of terrorist organizations which reject our model of civilized globalization but – paradoxically – use the technological instruments of our Information Society in order to promote their own fanatical and violent world-view. It is evident that – despite the development of the language industries and the global development of technology according to Zampolli's vision, and despite the work done in many cooperative European initiatives such as EAGLES, ISLE, FlaReNet and Meta-Net, as well as long-term regular international conferences in the field such as LREC or LTC – a long road still lies ahead of us if we are to put Roukens' vision of civilized globalization into practice. In other terms, there is still a lot to do for the LTC community.

***

In this book the reader will find a selection of 31 revised and in most cases substantially extended and updated versions of papers presented at the 7th Language and Technology Conference in 2015. The reviewing process was done by an international jury composed of the program committee members or experts nominated by them. The selection was made among 108 contributions presented at the conference and represents basically the preferences of the reviewers. Finally, the 82 authors of the selected contributions represent research institutions from the following countries: Czech Republic, Canada, France, Germany, Hungary, India, Japan, Nigeria, Poland, Romania, Serbia, Slovakia, Turkey and UK.

What are the presented papers about?

The papers selected in this volume belong to various fields of Human Language Technologies and illustrate a large thematic coverage of the LTC conferences. To make the presentation of

the papers possibly transparent we have "structured" them into 11 chapters. These are:

1. Speech Processing (4 papers)
2. Multiword Expressions (2)
3. Parsing (1)
4. Language Resources and Tools (4)
5. Ontologies and Wordnets (3)
6. Machine Translation (2)
7. IR/IE (Information and Data Extraction) (3)
8.Text Engineering and Processing (3)
9. Applications (2)
10. Emotions-Decisions-Opinions (EDO) (3)
11. Less-Resourced-Languages (LRL) (4)

Clustering the articles into chapters is approximate as many papers addressed more than one thematic area. The ordering of the chapters has not any "deep" significance; it roughly approximates the order in which humans proceed in natural language production and processing: starting with (spoken) speech analysis, through lexical and morphological analysis, (syntactic) parsing, etc. Within chapters we order the contributions in the alphabetic order according to the name of the first author.

Following this thematic order, we start this volume with the **Speech Processing** chapter containing four contributions. In the paper "Intelligent Speech Features Mining For Robust Synthesis System Evaluation" the authors (Moses E. Ekpenyong, Udoinyang G. Inyang and Victor E.Ekong) present their work on Deep Neural Networks (DNNs) applied to the evaluation of speech synthesis systems. The aim of the paper "Neural Networks Revisited for Proper Name Retrieval from Diachronic Documents" (Irina Illina and Dominique Fohr) is related to the application of neural networks in Out-Of-Vocabulary (OOV) proper names retrieval and the vocabulary extension of a speech recognition system. The third contribution concerning speech, "Cross-Lingual Adaptation of Broadcast Transcription System to Polish Language Using Public Data Sources" (Jan Nouza, Petr Červa and Radek Šafařík), presents methods and procedures that have been used to adapt to Polish a large vocabulary continuous speech recognition system (LVCSR) applicable to automatic broadcast transcription. The research on speech segmentation and recognition is the main issue of the last text "Automatic Subtitling System for Transcription, Archiving and Indexing of Slovak Audiovisual Recordings" (Ján Staš, Peter Viszlay, Martin Lojka, Tomáš Koctúr, Daniel Hládek and Jozef Juhár).

The **Multiword Expressions** chapter contains two papers. The first one on, "SEJF - a Grammatical Lexicon of Polish Multi-Word Expressions" (Monika Czerepowicka and Agata Savary), presents a lexical resource of Polish nominal, adjectival and adverbial multi-word expressions. The second, "Lemmatization of Multi-Word Entity Names for Polish Language Using Rules Automatically Generated Based on the Corpus Analysis" (Jacek Małyszko, Witold Abramowicz, Agata Filipowska, and Tomasz Wagner) is about the automatic corpus-based lemmatization of Multi-Word Units for highly inflective languages.

The **Parsing** chapter contains a single paper: "Parsing of Polish in Graph Database Environment" (Jan Posiadała, Hubert Czaja, Eliza Szczechla, and Paweł Susicki). The paper presents a rule-based syntactic parsing system for the Polish language using the Langusta natural language processing environment embedded in a graph database.

The **Language Resources and Tools** part is composed of four papers. The contribution "RetroC -- A Corpus for Evaluating Temporal Classifiers" (Filip Graliński and Piotr Wierzchoń) presents a corpus for training and evaluating systems for text dating. Authors of

the second article in this section, “Reinvestigating the Classification Approach to the Article and Preposition Error Correction” (Roman Grundkiewicz and Marcin Junczys-Dowmunt), reinvestigate the classifier-based approach to article and preposition error correction and claim that state-of-the-art results can be achieved without “(almost) no Linguistic Knowledge”. The next paper, “Binary Classification Algorithms for the Detection of Sparse Word Forms in New Indo-Aryan Languages” (Rafał Jaworski, Krzysztof Jassem and Krzysztof Stroński), presents an annotation tool used for semi-automatic tagging of New Indo-Aryan texts. The last paper of this chapter, “Multilingual Tokenization and Part-of-Speech Tagging. Lightweight versus heavyweight algorithms” (Tiberiu Boroș and Stefan Daniel Dumitrescu), proposes a framework for mobile devices which offers the essential state-of-the-art NLP tools.

The **Ontologies and Wordnets** chapter is composed of three papers. The paper “A Semantic Similarity Measurement Tool for WordNet-like Databases” (Marek Kubis) proposes a new framework for computing semantic similarity of words and concepts using WordNet-like databases. The next contribution, “Similarity Measure for Polish Short Texts Based on Wordnet-Enhanced Bag-of-Words Representation” (Maciej Piasecki and Anna Gut), presents a wordnet-based approach to semantic comparison of short texts. The last article of this chapter, “Methods of Linking Linguistic Resources for Semantic Role Labeling” (Balázs Indig, Márton Miháltz, and András Simonyi), is concerned with enriching the verb frame database of a Hungarian natural language parser by application of semantic resources as existing linguistic resources such as VerbNet and WordNet.

Two papers constitute the **Machine Translation** section. The authors of the first one, “A Quality Estimation System for Hungarian ” (Zijian Győző Yang, Andrea Dömötör, and László János Laki), present their approach to MT quality estimation in opposition to the existing automatic evaluation methods based on reference translations. The second paper, “Leveraging the Advantages of Associative Alignment Methods for PB-SMT Systems” (Baosong Yang, and Yves Lepage) discusses multi-processing-based new ideas in the statistical machine translation.

The **Information and Data Extraction** section contains three contributions. The first one, “Events Extractor for Polish Based on Semantics-Driven Extraction Templates” (Jolanta Cybulka and Jakub Dutkiewicz), presents a tool for identifying events in texts. The next paper, “Understanding Questions and Extracting Answers: Interactive Quiz
Game Application Design” (Volha Petukhova, Desmond Darma Putra, Alexandr Chernov, and Dietrich Klakow), presents a tool that extracts answers from unstructured Wikipedia texts. The last paper of the section, “Exploiting Wikipedia-based Information-rich Taxonomy for Extracting Location, Creator and Membership Related Information for ConceptNet
Expansion” (Marek Krawczyk, Rafał Rzepka and Kenji Araki), presents a method for extracting a number of semantic relations from Japanese Wikipedia XML dump files.

The next chapter is also composed of three contributions to **Text Engineering and Processing**. The chapter opens with the text “Lexical Analysis of Serbian with Conditional Random Fields and Large-Coverage Finite-State Resources” (Mathieu Constant, Cvetana Krstev and Duško Vitas) describing an approach to lexical tagging of Serbian texts combining three fundamental NPL instruments: part-of-speech tagging, compound and named entity recognition. What follows is the paper “Improving Chunker Performance Using a Web-based Semi-automatic Training Data Analysis Tool” (István Endrédy) focusing on issues related to noun phrase extraction from texts. The third contribution, “A Connectionist Model of Reading with Error Correction Properties” (Max Raphael Sobroza Marques, Xiaoran Jiang, Olivier Dufor, Claude Berrou and Deok-Hee Kim-Dufor), addresses a connectionist model of written word recognition with correction properties using associative memories.

Two papers are placed in the **Applications in Language Learning** chapter. The first one, "The Automatic Generation of Nonwords for Lexical Recognition Tests" (Osama Hamed and Torsten Zesch), proposes an automatic generation of tests for vocabulary proficiency of foreign language students. The second one, "Teaching Words in Context: Code-Switching Method for English and Japanese Vocabulary Acquisition Systems" (Michał Mazur, Rafał Rzepka and Kenji Araki), presents a system for computer assisted vocabulary learning using a code-switching method, in application to teaching Japanese vocabulary.

Contributions particularly concerned with linguistic expressions of **Emotions, Decisions and Opinions** were presented at the LTC within the EDO Workshop, integrated with the conference in form of a special track. Three papers have been selected for inclusion in this volume. The contribution "Automatic Extraction of Harmful Sentence Patterns with Application in Cyberbullying Detection" (Michał Ptaszyński, Fumito Masui, Yasutomo Kimura, Rafał Rzepka and Kenji Araki) describes a method for the automatic detection of malicious contents on the Internet. The second paper, "Automatic Extraction of Harmful Sentence Patterns with Application in Cyberbullying Detection" (Antoni Sobkowicz and Marek Kozłowski), is a comparative study of dictionary-based versus machine-learning-based methods in sentiment identification in web discussion texts. Finally, the paper "Saturation Tests in Application to Validation of Opinion Corpora: A Tool for Corpora Processing" (Zygmunt Vetulani, Marta Witkowska, Suleyman Menken and Umut Canolat) contributes to the discussion on corpora creation and validation issues with special attention payed to opinion corpora.

Less-Resourced Languages are considered of special interest for the LTC community and since 2009 the LRL workshop constitutes an integral part of the conference. In this volume, the LRL workshop is represented by four papers. The paper "Issues and Challenges in Developing Statistical POS Taggers for Sambalpuri" (Pitambar Behera, Atul Kr. Ojha, and Girish Nath Jha), reports on corpus collection and POS-annotation for Sambalpuri – a less resourced language spoken by some 0.5 million people only and with a small number of written text available for NLP research. The next one, "Cross-linguistic Projection for French-Vietnamese Named Entity Translation " (Ngoc Tan Le and Fatiha Sadat), faces the problem of named entity machine translation for the Vietnamese-French language pair. The third article, "National Language Technologies Portals for LRLs: A Case Study" (Delyth Prys and Dewi Bryn Jones) presents the initiative of a new Welsh National Language Technologies Portal as a *"resource for researchers, developers in the ICT and digital media spheres, open source enthusiasts and code clubs who may have limited understanding of language technologies but which nevertheless have a need for incorporating linguistic data and capabilities into their own projects, products, processes and services in order to better serve their wider LRL community"*. The paper "Challenges for and Perspectives on the Malagasy Language in the Digital Age" by Joro Ny Aina Ranaivoarison is the last of this section – as well as of the whole volume – and reports on the ongoing construction of an NLP dictionary of simple words (nouns, adjectives, adverbs, grammatical words) by using conventional dictionaries of the Malagasy language.

We wish you all an interesting reading.

February 2018

Zygmunt Vetulani
Joseph Mariani