

Preface

In the technologically advanced world, the Information Society Age has become reality. As we have observed on other occasions, this results from the transformation of the human technological environment of the twenties century into one that is not only rich in information but also an interactive environment. The growing role of language technologies, as predicted by Antonio Zampolli as early as the 1960s, is clear to those living in the Information Society Age, at least to those who can say “...way back then, when there was no Internet...”¹. This transformation is a product (or rather, a by-product) of long-term political and economic globalization.

Today’s globalization process is not the first in the history of humankind. Ancient empires and the Christian civilization that succeeded the Roman Empire serve well as earlier examples. In both cases, there was extensive exchange of ideas and codification of social norms as written laws and regulations. There was general acceptance of Latin as a lingua franca in those times. For many centuries, especially during the Middle Ages, Latin became the medium of communication for intellectual elites in different countries. This contributed to rapid technological development and to a new social stratification based on access to accumulated knowledge. Social and political revolutions between the eighteenth and twentieth centuries brought new ideas of liberty, equality, fraternity, and solidarity into play. These ideas sought general acceptance through a complex and often turbulent process that endures to this day. Nowadays, English has taken the place of Latin to facilitate the cross-border exchange of ideas and knowledge, but expectations have meanwhile increased: it is now commonly accepted in most countries that information and cultural values should be a part of a common heritage, shared and freely accessible to every human being. An ambitious programme of this kind can only be achieved by the use of appropriate technologies.

Language technologies, whose development is expected to speed up the process and to make the information circuit even more efficient and human-friendly, are in the mainstream of these global phenomena. For the same reasons as “ancient” global technologies (such as transport infrastructure, telecommunications, or classic media), language technologies need to be widely accessible, global, robust, and complete.² These requirements create new challenges. One challenging task is to fill the gaps in existing language resources and technologies. Gaps exist for most (if not all) languages, including English, because of variable technological development resulting from local political and economic constraints. From the first days of the LTC conferences, the organizers and the Program Committee have paid special attention to

¹ Such a phrase was recently pronounced by one of Zygmunt Vetulani’s students (20 years old) during a class on Human-Computer Communication.

² The META-Share initiative of 2011 redefined the requirements with respect to language technologies and resources in terms of *openness*, *distribution*, *interoperability*, and *security* (cf. <http://www.meta-net.eu>, access Nov. 12, 2011).

ensuring that the message has been communicated to the LTC participants, and several authors are in fact contributing to the discussion on these issues.

Reducing these gaps is a major challenge. In order to address this topic, a specific workshop on less-resourced languages was organized for the second time at the conference. This time, it was entitled “Less-Resourced Languages: Addressing the Gaps in Language Resources and Technologies”, and was organized by Khalid Choukri, Joseph Mariani, and Zygmunt Vetulani, with the support of ELRA, FLReNet, and META-NET. Closing existing gaps in language resources (LR) and technologies (LT) is a challenging task not only for less-resourced languages (LRL) but also for technologically more advanced ones. However, the needs of less-resourced languages are worth considering specifically in order to reduce imbalances between languages. The European Commission supports a large number of initiatives whose objectives include producing inventories and facilitating the sharing of language technologies and resources, including projects conducted within FLReNet, ELRA, META-NET, CESAR, METANET4U, or META-NORD such as LRE Map, Program Surveys, Language Matrices, Language Gaps, Language White Papers, and META-SHARE infrastructure. These are now available and contribute to better understanding the current landscape and devising possible solutions for each individual language and technology, including availability, quality, maturity, sustainability, and economic context. Various approaches have been proposed to address this challenge, either as monolingual processing addressing Basque, Luxemburgish, Quechua, or the Magahi Indian language, as bilingual processing for the Chinese–Japanese and Basque–Irish language pairs, and through multilingual processing, with specific focus on the 22 official Indian languages, as well as many other languages (French, English, Spanish, Vietnamese, Khmer, Pinyin Chinese, Ga, and Kisaninya). A great deal of effort is nowadays devoted to the open sharing of data, including sharing the methodology of LR creation and the collaborative definition of metadata and data formats. The cross-language portability of models, technologies, methodologies, and approaches is an important dimension to take into account at the design stage. Adopting a generic multilingual approach should therefore be encouraged. Actions should also be taken in the short term, as the gaps for less-resourced language may even get wider. A coherent action plan should be developed, based on several key factors: an open source approach, extensive standardization, the reuse of language foundations, tools and applications, and incremental design and development.

This volume contains the revised and in many cases substantially extended versions of 44 selected papers presented at the 5th Language and Technology Conference. This selection was made from the 111 high-quality contributions written by 253 authors accepted for presentation at the conference by an international jury using blind reviews. These evaluations were in general taken into account in the selection process for this volume, although there was some subjective selection. In a small number of cases, interesting papers presenting partial or preliminary results of ongoing research and development were not considered appropriate for publication in this volume, and it is hoped that more complete results will be available soon.

The selection of revised papers illustrates well the complexity and diversity of the field of Human Language Technologies. The papers collected in this volume report on

many man-years of hard work by 119 authors (and their teams) representing research institutions from 22 countries³: Australia, Bulgaria, Croatia, the Czech Republic, France, Germany, Hungary, India, Iran, Ireland, Italy, Japan, Malta, Nigeria, Norway, Poland, Serbia, Slovakia, Spain, Sweden, Switzerland, and the United Kingdom⁴.

The papers have been grouped into nine thematic chapters. Clustering the articles was a difficult task as in most cases the contributions addressed more than one thematic area, so our decisions should be considered as an approximation. In particular, the allocation of papers to chapters may not correspond to their allocation to LTC thematic sessions or to the authors' preferences. The nine chapters are:

1. Speech (3)
2. Parsing (3)
3. Computational Semantics (6)
4. Text Analysis (8)
5. Text Annotation (4)
6. Language Resources: General Issues (2)
7. Language Resources: Ontologies and Wordnets (7)
8. Machine Translation (4)
9. Problems Concerning Less Resourced Languages (7)

The ordering of the first four chapters follows the natural order that humans use to understand NL-messages: speech, morphology, syntax and semantics, text analysis. The next chapter focuses on text annotation and the following two on resources. Then comes the chapter on machine translation (MT), historically the first well-defined area of language engineering directly addressing applications of general interest. We close the volume with a large chapter (eight papers) that explicitly address the particular problems of less-resourced languages, traditionally an area of key interest at the Language and Technology conferences. We do not apply any specific ordering of contributions within chapters. Papers are presented in alphabetical order by the first author's family name.

The first chapter, Speech, contains three contributions. The first presents an implementation of two text-to-speech (TTS) systems for Ibibio, a language used in southeastern Nigeria, using a statistical-parametric approach (Ekpenyong, Udoh, Udosen, Urua)⁵. The next paper describes the design, development and evaluation of the Slovak dictation system for the APD judicial domain (Rusko, Juhár, Trnka, Staš, Darjaa, Hládek, Sabo, Pleva, Ritomský, Lojka). It is followed by a paper offering a further contribution to the intonational modelling of backchannels in Italian, useful for improving naturalness in voice-based dialogue systems for this language (Savino).

Three contributions appear in the Parsing chapter. We open with a presentation of a technique of combining multiple data-driven dependency parsers for parsing Arabic

³ Against 253 authors from 35 countries participating in the LTC 2011.

⁴ The language coverage is larger than may be inferred from the list of 22 countries. For example, languages such as Luxembourgish or Sanskrit do not correspond to the affiliations of the authors of some of the papers.

⁵ Moses Ekpenyong (University of Uyo, Nigeria) won a Best Student Paper LTC Award

(Alabbas, Ramsay). The next article presents a parsing algorithm to combine coordinated syntactic structures partially parsed using a coordination-free grammar (Dufour-Lussier, Guillaume, Perrier). In the last contribution of the chapter, an automaton-based lexical disambiguation process for lexicalized tree-adjointing grammar (Gardent, Parmentier, Perrier, Schmitz) is presented.

Various issues of Computational Semantics are of primary interest in the next chapter, which includes six contributions. It opens with a contribution to the automatic resolution of lexical anaphors in Sanskrit, the reference language for the Indo-Aryan family, with the help of POS tagging (Gopal, Jha). The second paper presents a multi-pass graph labeling approach for unsupervised co-reference resolution solved by relaxation labeling (Moosavi, Ghassem-Sani). In the third paper, the author presents the construction of the set of formal spatio-temporal relations expressed in the XCDC model that are conceptualized in a natural language (Osiński). Sentiment analysis is the focus of the fourth paper, where the authors show how a simple technique can improve classification by normalizing term weights in the basic bag-of-words method (Pak, Paroubek, Fraisse, Francopoulo). The next paper describes work on textually recording non-linear elements of Polish Sign Language utterances (Romaniuk, Suszczańska, Szmaj). The last article of this chapter presents a method for resolving certain semantic ambiguities by means of a constraint calculus (RCC-5) and an algorithm for spatial reasoning (Walas, Jassem).

Eight papers appear in the Text Analysis chapter, opening with a paper on a complex system of natural language processing that performs extraction of basic facts as result of morphological, syntactic and semantic analysis (Baisa, Kovář). The second paper deals with text normalization applied to a historical dialect of German (Early New High German between 1350 and 1650) by mapping historical word forms to modern ones (Bollmann, Petran, Dipper). A rule-based method to capture the tense, aspect, and mood (TAM) features of Hindi verb groups was investigated by the authors of the third paper (Choudhary, Pandey, Jha). The next paper presents research on reference resolution and personal name recognition in German Alpine texts (Ebling, Sennrich, Klapper). The fifth paper is a contribution to supervised machine learning techniques applied to topic tracking problems (Fukumoto, Suzuki, Yamamoto). The next paper will be of interest to those concerned with the recognition of the textual extent of temporal expressions (Mazur, Dale). The seventh paper presents results on applying text mining techniques to named entity recognition (Nouvel, Antoine, Friburger). The last contribution is about the identification of lexical bundles in school textbooks (Ribeck, Borin).

Four papers form the Text Annotation chapter. The first presents a framework for dialogue act annotation performed semi-automatically by using statistical models trained on previously annotated dialogues (Ghigi, Martinez-Hinarejos, Benedi). In the second, the authors propose a method to calculate confidence measures for an automatic dialogue annotation model (tested on a task-oriented human-computer corpus of railway information; Martinez-Hinarejos, Tamarit, Benedi). The third contribution presents a version of the morphosyntactically annotated Serbian translation of Orwell's *1984* and the resources used for annotation (Krstev, Vitas, Trtovac). The last paper of the chapter reports on video analysis algorithms used for the annotation of video recordings in language diversity preservation, considered an important part of

the cultural heritage of humanity (Lenkiewicz, Drude, Lenkiewicz, Gerbe, Manseri, Schreer, Schwenninger, Bardeli).

The chapter entitled “Language Resources: General Issues” contains two articles. The first is a position paper about crowdsourced microworking systems in frequent use for the collection and processing of language resources (Fort, Adda, Sagot, Mariani, Couillaud). The second paper discusses the relationship between language resources conceived from a local perspective and a shared framework conceived from a global perspective that supplies such resources for local re-use or enhancement (Rosner, Attard, Thompson, Gatt, Ananiadou).

The next chapter entitled “Language Resources: Ontologies and Wordnets” contains seven contributions. The purpose of the first paper is to show how to add sense descriptions to GermaNet from Wiktionary automatically (Henrich, Hinrichs, Vodolazova). The second article describes a data manipulation language designed for WordNet-like lexical databases (Kubis)⁶. The third is on transforming a subject heading language used in Polish library catalogues (KABA) into a fully machine-readable thesaurus (Mazurek, Sielski, Walkowska, Werla). The fourth contribution is about enhancing the functionality of tagging systems in the field of history and culture by introducing WordNet-like ontologies (Marciniak). The author of the next paper addresses issues connected with natural language based ontologies and proposes an ontology infrastructure (IMAGACT) as a way to fill the current gap in action ontologies (Moneglia). The next is a presentation of an automatic, language-independent approach to extending an existing wordnet by recycling freely available bilingual resources (Sagot, Fišer). The closing article of this chapter presents recent advances in the long-term project “PolNet – Polish WordNet” since the first public release of PolNet 1.0 presented at LTC 2011 (Vetulani).

The next chapter of four papers contribute directly to Machine Translation. In the first one the authors describe an approach to improve the performance of sampling-based multilingual alignment on translation tasks by investigating the distribution of n-grams in the translation tables (Luo, Lardilleux, Lepage). The second proposes a basic research for example-based machine translation consisting in a study of analogies between chunks of text in 11 European languages (Takeya, Lepage). The third paper presents a comparative study of Corpus-Based MT paradigms: statistical (SMT), example-based (EBMT), and a hybrid (EBMT-SMT) one for the German–Romanian language pair (Gavrila, Elita). In the last paper of the chapter the authors demonstrate that the style of the training corpus may influence the quality of the translation output in statistical machine translation systems (Gavrila, Vertan).

Finally, the volume ends with seven papers classified as contributions to problems specifically addressing less-resourced languages. The first contribution of the LRL chapter is concerned with the analysis of observations collected in the Language White Papers of the CESAR project. The objective of this study is to fix gaps in Language resources, tools and services for the six languages included in the project, i.e., Bulgarian, Croatian, Hungarian, Polish, Serbian, and Slovak (Tadić, Váradi, Garabík, Koeva, Ogrodniczuk, Vitas). The next two papers each address a particular

⁶ Marek Kubis (Adam Mickiewicz University, Poland) won a Best Student Paper LTC Award.

language: Luxembourgish and Magahi. In the first one, the first LVASR (large vocabulary automatic speech recognition) system for Luxembourgish is presented (Adda-Decker, Lamel, Adda, Lavergne). The second describes the first attempt to develop an annotated corpus of Magahi, an Indo-Aryan language spoken by 14 million people, but until now without language resources and with few written texts (Kumar, Lahiri, Alok). A special case of bilingual processing is the focus of the next paper, where adapting to Irish a multi-choice quiz generation system originally developed for Basque and English is discussed from the LRL's point of view (Maritxalar, Ui Donnchadha, Foster, Ward). Finally, the last three papers are about issues pertaining to several languages of this group. The first presents a generic approach to text normalization, important for building multipurpose multilingual text corpora involving LRLs (Bigi). The second paper describes parallel corpora created simultaneously in 12 major Indian languages including English (Choudhary, Jha)⁷. The last paper in the volume presents a strategy for the incremental construction of deep parsing grammars on the basis of interlinear glossed text data, available for languages with little or no digital resources (Hellan, Beermann).

February 2014

Zygmunt Vetulani
Joseph Mariani

⁷ Narayan Choudhary (Jawaharlal Nehru University, India) won a Best Student Paper LTC Award.

Human Language Technology Challenges for Computer
Science and Linguistics

5th Language and Technology Conference, LTC 2011,
Poznań, Poland, November 25--27, 2011, Revised
Selected Papers

Vetulani, Z.; Mariani, J. (Eds.)

2014, XVI, 550 p. 109 illus., Softcover

ISBN: 978-3-319-08957-7