Special Section on Corpora

Guest editors: NICHOLAS OSTLER and ANTONIO ZAMPOLLI

Introduction to Part Two

ANTONIO ZAMPOLLI
University of Pisa, Pisa

Abstract

The articles presented in this second part of the Special Section on Corpora discuss the following issues related to the construction and use of large-scale linguistic resources (written and spoken corpora and lexica):

—quantitative methods in processing natural languages:
—extraction of knowledge of lexical and textual projects, computational and traditional lexicography, and statistical and rule-based approaches;
—the use of linguistic resources in assessment and evaluation;
—the role of standards in creating and sharing multifunctional linguistic resources.

The second part of this Special Section on Corpora (the first has been published in volume 8, number 4) includes four additional papers relevant to the NERC International Workshop held in Pisa in January 1992.

It is well known that the success of the use of corpus-based quantitative models in the field of speech processing has been a key factor in reviving the interest of the research and development Natural Language Processing (NLP) communities in the collection and use of language corpora. The combination of quantitative methods with rule-based methods in language analysis is today one of the most challenging topics in the NLP field.

It should be noted that the speech processing work is enriching the variety of language corpora with new corpus types. In particular, as A. Fourcin and D. Gibbon state in their article 'Spoken Language Assessment in the European Context', 'unlike written language corpora, which are typically collated from existing texts and are associated with technical, legal, and logistic problems arising from this fact, spoken language corpora collected in the speech processing paradigm are typically custom-designed for a particular task, highly controlled, and associated with specific technical, experimental and logistic design and production problems'. This article presents the work and results of the ESPRIT 'Speech Assessment Methods' (SAM) project, which concerns the establishment of database, speech, and language descriptive methods, and quantitative tools for the assessment both of speech recognizers and speech synthesis systems, with the general aim of contributing to introduce robust techniques for real-time multilingual operations. We should also note the outstanding pioneering contribution of the SAM project to the promotion of commonly agreed standards for linguistic resources. Common methods of linguistic standardization have been evolved and a common computer compatible phonemic notation (SAMPA) has been introduced for European languages. The article outlines a structured perspective for standards and resource-oriented themes in future speech and spoken language work.

Lexicographers have traditionally based their work on the analysis of citations extracted from large textual corpora. Today, also the NLP community is increasingly aware of the need, for realistic applications, of large lexica based on the evidence of real language use provided by large corpora analysis. The study of appropriate methods to extract the relevant knowledge from texts is an outstanding issue in NLP. The article of R. Bindi, N. Calzolari, M. Monachini, V. Pirrelli, and A. Zampolli, 'Corpora and Computational Lexica: Integration of Different Methodologies of Lexical Knowledge Acquisition', stresses the necessity of a methodological framework for the convergence of lexical and textual projects, computational and traditional lexicography, and statistical and rule-based approaches. It presents an attempt to integrate different techniques and various perspectives on lexical knowledge acquisition from text corpora. The authors use three distinct methodologies to handle text data:

(i) Simple and traditional stochastic techniques working on pairs of words.
(ii) A lexicographic approach, aiming at a formal description of sense disambiguation in terms of rules.
(iii) More complex and sophisticated statistical methods which should allow a new perspective on the problem of sense disambiguation. The three approaches are complementary to each other and can be contextually used.

Also J. Cowie, T. Dunning, L. Guthrie, and Y. Wilks, in their article 'Text Processing Using Multilingual Resources at the Computing Research Laboratory', discuss, through the survey of the work on multilingual text processing of their Las Cruces Laboratory, various aspects of the use of large-scale linguistic resources (corpora, machine-readable dictionaries, etc.) for a range of information extraction tasks (IE) and in particular:

—The automation, to the greatest degree possible, of the gathering of the linguistic resources needed for IE.
—The role of automatically constructed lexicons (from machine readable dictionaries and corpora) in the IE task.

Correspondence: Antonio Zampolli, Istituto di Linguistica Computazionale, CNR, Via della Faggiola 32, 56100 Pisa, Italy.

—The role of active semantic and knowledge structures, seeking instantiation in a text, in the IE process.
—The beneficial role of symbolic-statistical hybrid systems that emphasize the benefits of both methodologies within IE.

Past experience has clearly shown that the creation of adequate large-scale linguistic resources is a costly enterprise, impossible to be carried out by one single organization. Duplications must be avoided as far as possible. Not only the financial, but also the human resources, possessing adequate skills and know-how, are limited. Until now, it has been the usual practice for each project to construct its own *ad-hoc* lexicon, for a specific research and application, restarting from scratch even within the same company and research team. In this context, the 'reusability of linguistic resources' has become a key concern in the field of computational linguistics and its applications. This expression, which appears more and more frequently, in the definition of the objectives of national and international projects, refers to two major complementary issues. The first one concerns the ability to reuse existing linguistic resources by extracting or converting their data for incorporation in a variety of different language-processing modules. The various types of existing dictionaries available in machine-readable form have already proved to be rich and valuable sources of information, in particular by the research of ACQUILEX, a project funded by the European Community within the framework of the ESPRIT Basic Research Program. The second issue concerns the need to design new large linguistic resources so that they have the property of being multifunctional, i.e. capable of serving, through appropriate interfaces, a wide variety of present and future research and applications. A particular crucial and controversial problem is to define whether, in order to be reusable and multifunctional, linguistic resources should, and could, also have the property of being polytheoretical, i.e. usable in different linguistic theory frameworks. As a matter of fact, differences among the requirements of the lexical information of NLP systems could be determined not only by the intended applications, but also by the specific linguistic theory on which the components of the system (parsers, generators, etc.) are explicitly or implicitly based.

Multifunctionality, reusability, shareability require the adoption of *de facto* standards in the construction of linguistic resources. The need for standards is today widely recognized in various frameworks. The last two papers present two projects, one, MULTILEX, sponsored in the framework of the ESPRIT programme of the European Commission, the second, GENELEX, supported in the framework of EUREKA, which are both contributing to the definition of standards for the representation and use of lexical knowledge.

In the article: 'Use and Importance of Standard in Electronic Dictionaries: the Compilation Approach for Lexical Resources', H. Khatchadourian and N. Modiano present and discuss, from a software engineering point of view, the idea of compiling application-specific lexica on the basis of a standardized lexical database. 'In contrast to work focusing on either one of the two above mentioned aspects of reusability, the compilation approach is intended to integrate both aspects, and to optimally support the design of natural language products and deals with the entire life-cycle of the lingware engineering related to electronic dictionaries.'

M. H. Antoni-Lay, G. Francopoulo, and L. Zaysser, in the article 'A Generic Model for Reusable Lexicons: the Genelex Project', present the GENELEX activities, whose main goals are:

—To define a generic model for lexicons.
—To design and develop software tools for lexicon management.
—To apply the model and the tools to dictionaries.
—To build full-scale electronic dictionaries.

'The requirements a generic model most satisfy in order to be "theory-welcoming" and to have a broad linguistic coverage are discussed in this article, taking syntactic data as examples'.

The GENELEX model 'offers means to encode syntactic information according to different lexicographic points of view, and to unify pieces of syntactic description that originate in different theories'.

The efforts of projects like ACQUILEX, MULTILEX, GENELEX, NERC, and SAM has led the CEC Community to launch the EAGLES project, whose aim is to work towards the establishment of *de facto* standards for lexica, corpora, formalisms, evaluation and assessment, and speech data, based on the consensus of academic and industrial European actors, and in particular of the representatives of the major European international projects.

Concluding remarks

The crucial factor that made it possible to propose the creation of adequate textual corpora and large computational lexica as major research and development areas in which computational linguistics should join the efforts of linguistics, lexicography, psycholinguistics, various types of humanities, etc., was the progressive diffusion, in the second half of the 1980s, of the so-called 'language industry paradigm'.

The term 'language industries', launched at the 1986 Tours Conference organized by the Council of Europe, includes both computer assistance to traditional applied linguistics professions (e.g. among others, lexicography, translation, langague teaching) and the development of computational systems based on NLP (as required, for example, in natural language interfaces, speech analysis and synthesis, automatic indexing and abstracting, office automation, machine translation, and, more generally, support for communication). Language industry products, ranging from spelling checkers to information retrieval to machine translation, require robust NLP components capable of dealing with 'real texts'.

Textual corpora have been recognized as the essential sources of data for the description of the real uses of natural languages in the various communicative contexts. It has also been recognized that, for real world

applications, NLP systems must be able to deal with tens and hundreds of thousands of lexical items. Consequently, the development of large textual and lexical resources has emerged as one of the most urgent tasks in the language industry framework. A major turning-point was the workshop 'On automating the lexicon', held in Grosseto (near Pisa) in May 1986.

This workshop was organized by ILC, Pisa in strict cooperation with Don Walker, and was also the starting-point for several cooperative initiatives between researchers working in the paradigms of NLP and literary and linguistic computing (LLC).

I feel it appropriate to recall here the fundamental role of Don Walker, who recently passed away, leaving a great void in our community. His dedicated, enthusiastic and far-sighted actions were instrumental in promoting cooperative efforts between ACL, ALLC, and ACH: for example, the joint sponsorship of TEI (Text Encoding Initiative) and various panels, conferences, and projects.

The 1992 Pisa workshop on corpora, in which Don participated, has confirmed that the construction and use of large-scale linguistic resources could be a strong stimulus to foster the cooperation between NLP and LLC.

Literary and linguistic computing has always been interested in the process of large real texts, but the computational treatment has been performed on units identified, mainly if not exclusively, at the graphical level.

However, several operations on the texts, which form an integral part of various scholarly humanistic activities, are based on the identification, in the text, of linguistic units at various levels, both as direct objects of linguistic, philological, literary research, and as referential units representing factual information.

The intrinsic complexity of the analysis, and the time required to perform it, are very high. Therefore, LLC should consider the possibility of cooperating with NLP in the construction of tools for automating, at least in part, the operations of analysis.

I would call the attention, in particular, on two major categories of tools:

(i) Robust parsers, supported by large computational lexica, conceived for identifying, in real texts, linguistic units, at various levels of analysis: syllabic, metrical, syntagmatic patterns; lemmata; parts of speech; phrases; verbal arguments; superficial sentential structures; etc.

(ii) 'Intelligent' access tools which, through the consultation of various kinds of knowledge sources, assist the researcher in the interaction with the texts. For example, appropriately structured reference sources, such as encyclopaedias, dictionaries, semantic taxonomies, etc., can make explicit, and eventually complement, the linguistic and conceptual researchers' knowledges, in such a way that they can be used by the programs to assist the researches in text browsing.

Summing up, both NLP and LLC are led by various factors, and in particular by the framework created by the expansion of the 'information society', to consider the creation of tools and language resources, for the processing of large real texts, as a major task in their present state of development. The basic knowledges required are in large part the same. It is therefore important that the information encoded can be reused in both fields through appropriate interfaces. Cooperation must be promoted, in order to combine the efforts and specific know-how of the researchers of the two fields, which are, in several aspects, complementary. For example, NLP could contribute grammatical formalisms, taggers, and parser models; LLC has developed knowledge and methods for corpora collection and treatment, statistical linguistic analysis, sublanguage description and identification.

Large lexical and textual knowledge bases are considered to be precompetitive resources, which, therefore, must be promoted in the public domain. Multilingualism is a central aspect of language industries. 'Informatization' has been indicated as a key element for the conservation of the vehicular function of a language, and therefore for its preservation, which is, in turn, an important factor for the preservation of the national cultural identity. National and supranational authorities are recognizing that the growing flow of multilingual information, in the worldwide economic system and in the telecommunication and 'information society', puts an obvious pressure for the development of new products, based on automatic processing of natural languages, which are still the principal vehicles for producing and storing information.