Problems of computational Lexicography

Thierry FONTENELLE

WALKER (Donald E.), ZAMPOLLI (Antonio) and CALZOLARI (Nicoletta): 1995, *Automating the Lexicon: Research and Practice in a Multilingual Environment* (Oxford: Oxford University Press), 1995, 413 pp. Price £50. ISBN 0-19-823950-5.

This book comprises the long-awaited proceedings of a workshop entitled *Automating the Lexicon: Research and Practice in a Multilingual Environment* which was held in Marina di Grosseto (Italy) in May 1986. The Grosseto Workshop marked a turning- point in what had come to be known as computational lexicography and lexicology and is usually considered as a point of departure for major research projects, summer schools and international conferences devoted to the creation and exploitation of machine-readable dictionaries (MRDs) and lexical databases (LDBs). The editors rightly point out that "although this book has 'historical' relevance, it addresses many issues that are still being debated today and that guide research and development efforts" (p. 1).

The Grosseto Workshop brought together a wide range of academic and industrial R&D groups, publishers and commercial firms interested in marketing lexical resources. The growing interest in the lexicon which had motivated this workshop was a new phenomenon at the time, mainly because the generative tradition of the 1960s and 1970s in linguistics had tended to neglect the role of the lexicon in the description of language. The advent of computer technology, however, now enabled linguists to test their hypotheses and intuitions and it had become apparent that the development of a whole range of natural language processing (NLP) systems required a lot more information about words than what was needed to parse a lexically simple sentence such as "John loves Mary". Developers were rapidly confronted with the so-called 'lexical acquisition bottleneck'. The problem was very simple: in order to develop a large-scale NLP system, *e.g.* a machine translation (MT) system or an information retrieval (IR) system, one needs to feed the lexical component of the system with the

✉ University of Liège; English Department; place Cockerill, 3; B-4000 Liège (Belgium).
Fax: +32 41 66 57 21 E-mail: u016601@vm1.ulg.ac.be

description of tens of thousands of lexical items. The provenance of these descriptions, however, is rather problematic. Should the developer hire a team of highly specialized (and expensive) lexicographers to code the whole lexicon from scratch or should other resources be tapped to reduce the costs and save time? The Grosseto Workshop definitely opted for the latter approach, *i.e.* exploiting existing lexical resources to build the lexicon of these NLP applications and recognizing that MRDs house a lot of NLP-relevant information the extraction of which can be partly automated. The workshop also gave rise to a number of recommendations which are summarized in the editors' introduction to the book. In particular, this summary stresses the need to clarify the copyright issues associated with the various lexical resources and to create and maintain registries of MRDs and text corpora, together with repositories of available material. The need to come up with well-defined terminological, linguistic and lexicographical conventions (standardization) was also alluded to, as was the need to study the interactions between various types of linguistic factors in a word sense assignment perspective. The introduction also includes a survey of major activities in the lexical area, a most welcome initiative given the profusion of research projects centered around the notion of 'reusability of lexical resources'.

The main bulk of the book consists of 11 papers, some of which have been slightly updated since their preparation.

In *Identifying the linguistic foundations for lexical research and dictionary design*, Richard Hudson questions a number of widely-held assumptions, viz. (a) that the lexicon is a distinct component of grammar, (b) that there are discrete lexical entries, and (c) that the 'lexico-grammar' contains only intralinguistic information. His contention is that linguists should develop theories which reflect the fact that there is no natural boundary between the lexicon and the grammar. The tendency for learner's dictionaries to include syntactic information in the form of more or less explicit grammatical codes is therefore seen as a positive improvement which should inspire 'mainstream' linguists. Along the same lines, Hudson argues that the lexicon should also contain encyclopedic information and that connections among lexical items should be made explicit, thereby moving towards a conception of the lexicon as a network or a relational database. Even if one does not necessarily share Hudson's bias towards his own linguistic formalism, viz. Word Grammar, the problems he addresses in his paper raise crucial questions and are still debated today.

In the second paper, entitled *Approaches to lexical semantic representation*, Beth Levin surveys efforts to represent semantic information and to come up with an adequate representation of the meaning of a word. This paper, which had been distributed in pre-print form shortly after the workshop, exercised a profound influence on later attempts to investigate the syntax-semantics interface in computational linguistics. Levin's in-depth study of a range of transitivity alternations (and more specifically of the causative/inchoative, a.k.a. ergative, alternation) called for the formulation of a lexical-semantic representation to make explicit the semantic relations between a verb and its arguments. The underlying assumption, which Levin further developed in a more recent book on *English Verb Classes and Alternations* (1993, the University of Chicago Press), is that the behaviour of a verb, particularly with respect to the expression and interpretation of its arguments, is to a large extent determined by its meaning. One immediately sees the implications this assumption can have for NLP, since the size of the lexicon can be dramatically reduced if one is able to make such generalizations about regularities in the expression of arguments.

In *Lexical information for parsing systems: Points of convergence and divergence*, Robert Ingria presents the various types of information that must be represented in a lexicon for parsing. Obviously, this paper reflects state-of-the-art systems in the early and mid-1980s, which means that many of the NLP systems described in it now belong to the history of computational linguistics. The basic problems it addresses are still relevant today, however, if one considers the various decisions developers have to make when constructing the lexical component of NLP systems (*e.g.* should word senses be conflated or distinguished? How should collocations and idioms be represented? Is it possible to share information among different systems?...).

Susanna Cumming then addresses the problem of *The lexicon in text generation*. Her goal is to describe what is (or was) being done in the field but also what a generation lexicon could do. After briefly examining what is word knowledge, she describes a number of generation systems and devotes several interesting sections to co-occurrence phenomena such as collocations (*i.e.* lexical restrictions on the modifiers of an item), idioms and selectional restrictions.

In *Dictionary systems for office practice*, Roy Byrd, the then head of the Lexical Systems group at *IBM* Yorktown Heights, describes the functions dictionary systems ought to provide to their users. The basic problems related to the construction of dictionary systems indeed depend on whether the dictionaries are for people (on-line dictionaries for direct use at the user's workstation) or for programs (applications supporting natural language interfaces). This paper mainly presents work done at *IBM* in the 80s, and is therefore perhaps less relevant today, although the strategies it advocates should be taken into account by the developers.

Jonathan Slocum and Martha Morgan's paper on *The role of dictionaries and machine-readable lexicons in translation* is especially concerned with bilingual dictionaries in an MT perspective. The numerous figures and diagrams are based on the METAL German-English MT system (now developed by *Siemens-Nixdorf*), which means that some of the positions held here would probably not be tenable in other, more recent systems based on unification formalisms. This chapter is most welcome, however, because the emphasis in the literature is traditionally laid on the description of monolingual problems, very much as if the bilingual perspective were less interesting or raised fewer theoretical issues.

Judy Kegl then tackles the problem of *Machine-readable dictionaries in education*, an area which has attracted less attention, although computational lexicographers are often using the computerized versions of learner's dictionaries, *i.e.* resources primarily intended for non-native speakers learning foreign languages. The perspective is here the creation of a number of exercises from WordNet, a large lexical network derived from a range of MRDs. Electronic dictionaries are viewed as educational tools used to teach people how to determine which of the multiple senses of a word is most appropriate.

Michael Lesk addresses another use of MRDs, namely *The use of dictionaries and thesauri in information retrieval* (or *Why use words to label ideas?*). The key issue here is how to construct a thesaurus for retrieval or automatic indexing. The final question Lesk asks in his conclusion is whether it is possible to imagine combining a dictionary and a thesaurus. In a way, this question is answered, positively, in Nicoletta Calzolari's paper on *Structure and access in an automated lexicon and related issues*, which discusses some general requirements for a lexical database (LDB) and describes the various types of lexical and semantic relations implemented in the Italian LDB developed at the University of Pisa.

Branimir Boguraev's contribution on *Machine-readable dictionaries and computational linguistics research* describes the rationale for the use of MRDs in computational linguistics. His survey of MRD-based research sheds light on the history of this relatively new field and the section devoted to the reliability of MRDs should be read by all those who embark on the exploitation of computerized dictionaries and hope to be able to retrieve information from them with 100 % accuracy.

The final paper, prepared by Susan Armstrong-Warwick, presents a *Survey of automated lexical resources in Europe*. She recognizes that "activities in the field have expanded so rapidly that only a sampling of the numerous new projects and products is included in this version" (p. 357). Some major projects such as the Collins Cobuild dictionary (Sinclair, ed., HarperCollins Publishers, 1987) are not mentioned in this survey and other projects which are described have evolved so much that the figures describing them are no longer up-to-date. I am thinking more particularly of the illustration, p. 390, of a Robert & Collins database entry which reflects the state of the Liège database in 1986 and which is now very far from what I have used, exploited and developed in my own doctoral dissertation on the topic! Despite this inevitable limitation, this type of survey is extremely useful insofar as it describes various types of computerized dictionaries and provides food for thought to researchers and developers who are faced with the gnawing question of deciding on the form and structure of their lexicons.

To conclude, *Automating the Lexicon* certainly contains papers that only have historical relevance because some of the data they present reflect the situation of computational lexicography in 1986. Many issues that are discussed in the book are still debated today, however, and some of the papers here can be considered as seminal contributions which exercised a profound influence upon the way dictionary data are used and exploited. The list of European and American projects given in the introduction testifies to the resurgence of interest in lexical matters over the past few years. The papers in this book have certainly contributed to it and should therefore be read by anyone who wishes to know more about computational lexicography and lexicology.