

## Lexicon Development for Speech and Language Processing

Frank Van Eynde and Dafydd Gibbon (editors)  
(University of Leuven and University of Bielefeld)

Dordrecht: Kluwer Academic  
Publishers (Text, speech and language  
technology series, edited by Nancy Ide  
and Jean Véronis, volume 12), 2000,  
xi+298 pp; hardbound, ISBN  
0-7923-6368-X, \$128.00, £79.00, €109.00

*Reviewed by*  
*Ken Litkowski*  
*CL Research*

As a computational lexicologist with little background in speech technology, I approached this book as an opportunity to gain a basic understanding of how lexicons are used in speech and to see whether semantically oriented lexicons from my work could bring something to this field. The first objective was met, but the second was not. This book arose from the 1997 Fifth European Summer School on Language and Speech Communication under the auspices of the European Language and Speech Network (ELSNET) and is the report of nine of the ten lectures. The book offers a “survey of methods and techniques for structuring, acquiring and maintaining lexical resources for speech and language processing” (p. ix).

As part of my first objective, I learned that speech technology is greatly hampered by limitations on the size of the lexicon that can be handled efficiently in real systems. The amount of data is overwhelmed by the need to use signal data in analog or digital form. The speech community has to resort to and develop many tricks for working with the prodigious amount of data. This book provides an appropriate overview of the complexity of the problems and will serve well as background reading in an introductory computational linguistics course. Each of the papers is well referenced, with the authors providing their (important) guides to further details.

The organization of the book is not optimal, however. It consists of an overview, two papers on lexical formalisms, two research papers, three database papers, and a final research paper. Logically, the database papers should be first, followed by the research papers, the lexical formalisms, and finally the overview. The quality of the papers and the editing is high; the contributors clearly worked to turn their lectures into readable form. I will present comments based on the logical order, rather than the physical order.

Christoph Draxler (“Speech databases,” Chapter 6) provides an overview of what kind of database technology is required in speech technology, albeit with little reference to the lexicon. This paper shows the impressive range of speech material, describing recording mechanisms and annotation levels. Silvia Quazza and Henk van den Heuvel (“The use of lexica in text-to-speech systems,” Chapter 7) outline practical steps for putting together the components of a text-to-speech system and for building the lexicon, showing quite well where the lexicon comes into play. This chapter has the most detailed information about what lexical information is used in speech systems and how. Martine Adda-Decker and Lori Lamel (“The use of lexica in automatic speech recognition,” Chapter 8) carry us through the design and development of the lexical information used in recognition, describing what is required of the lexicon for de-

coding speech. It is here that the size of the lexicon evinces the complexities and is likely the place where ingenious (semantic) solutions are most needed for recognizers to “understand” speech. These three papers provide the backbone of the book.

The three research papers represent examples of ongoing research; all are presented quite well, giving the reader a sense of active problems. Walter Daelemans and Gert Durieux (“Inductive lexica,” Chapter 4) describe their work on inducing regularities implicit in phonological lexical representations using “memory-based learning” machine learning techniques. The methods followed are clearly laid out, providing an introduction to the authors’ research, which may be followed via their references. The other two research papers do not have a close tie to speech. R. Harald Baayen, Robert Schreuder, and Richard Sproat (“Morphology in the mental lexicon: A computational model for visual word recognition,” Chapter 9) present investigations on the relative time of lexical processing for morphologically complex words; such research may eventually help in the design of computational lexicons for speech processing, via activation of potentially matching lexical candidates. Gregory Grefenstette, Anne Schiller, and Salah Ait-Mokhtar (“Recognizing Lexical Patterns in Text”, Chapter 5) provide an introduction to finite-state automata for recognizing compound noun patterns in building a lexicon; in addition to its introductory pedagogic value, the paper describes a fully developed system.

Gosse Bouma, Frank Van Eynde, and Dan Flickinger (“Constraint-based lexica,” Chapter 2) provide an introduction to the HPSG formalism. Although the presentation is clear, there is no tie to speech or to speech systems. Lynne Cahill, Julie Carson-Berndsen, and Gerald Gazdar (“Phonology-based lexical knowledge representation,” Chapter 3) present a tutorial on the DATR formalism, valuable in itself, but more importantly they demonstrate how this formalism can be used for phonetic or phonological representations. There are 34 “exercises” in the tutorial; however, these may not be as useful in the book as they would have been in the summer school.

The introduction and overview to the book (“Computational lexicography,” by Dafydd Gibbon), which needs to be read both first and again after reading the other chapters, brings together the papers in the volume under the author’s development of an “integrated lexical sign model.” This is a little forced, but useful for providing an overview of the lexicon in speech processing. However, there is a strong element of Unix hacking, without a comprehensive view of traditional computational lexicology and no reference to computational lexicography as may be practiced by dictionary publishers (e.g., corpus evidence).

As to the second objective stated above, the incorporation of semantics into speech seems a long way off. However, the dialogue can begin.

*Ken Litkowski* is a computational lexicologist with CL Research (<http://www.clres.com>). He is currently performing computational *lexicography* tasks for dictionary publishers and performing computational *lexicology* research on how computational lexicons from machine-readable dictionaries and thesauruses can be used in NLP applications, including word-sense disambiguation and question answering. Litkowski’s address is: CL Research, 9208 Gue Road, Damascus, MD 20872; e-mail: ken@clres.com.