Linguistic Databases

John Nerbonne (editor)
(University of Groningen)

Stanford: CSLI Publications (CSLI
lecture notes, number 77) (distributed
by Cambridge University Press), 1998,
xii+243 pp, hardbound, ISBN
1-57586-093-7, $64.95; paperbound,
ISBN 1-57586-092-9, $24.95

Reviewed by
Jörg Tiedemann
Uppsala University

Linguistic Databases is an edited collection of papers on the use of databases in linguistics. It comprises a selection of 12 contributions to the conference with the same title, which was held at the University of Groningen on 23–24 March 1995.

The need for data management tools in linguistics is evident. Although collections of linguistic data grew rapidly in the past, the development of suitable database structures and management systems is still in an early stage. The articles presented in the book introduce a variety of approaches to several kinds of applications in different fields of linguistics. They are generally based on existing encoding schemes and data management standards. The fields of study considered here include data-organization approaches to syntactic corpora and phonetic data, the management of theoretical linguistic data, such as syllable structures and nominal argument structures, applications to linguistic problems such as the simplification of texts, and the extension of existing systems and the interaction between them.

Two articles consider the management of "test suites" of linguistic phenomena. In the first article, by Stephan Oepen et al., linguistic examples of language-specific phenomena, such as complementation and agreement, were organized in a relational database structure that can be linked to a syntactic parser and grammar. The project emphasizes the development of a consistent annotation scheme and introduces two implementations based on standard tools and a public-domain C library, respectively, based on the commercial DBMS FoxPro (Microsoft). In the second article, the application of SGML for the annotation of linguistic test suites is discussed. The author, Martin Volk, shows the applicability of SGML for this task but also points out problems with redundancy and efficiency when using an SGML annotation scheme. A third article on syntactic data management introduces an approach that combines standard database query systems with SGML-encoded texts. A newly defined query language, SgmlQL, is described, which represents an extension of the standard query language for relational databases SQL. In this language, database queries can be formulated and applied to process hierarchic SGML structures (tree manipulation), for instance, in order to extract information. Furthermore, a freely available prototype was developed that implements a subset of this query language.

Four articles deal with phonetic data. Werner Deutsch et al. introduce an implementation of a database management system (S-Tools) for acoustic data. This system includes graphical editors and specialized tools for classifying and handling phonetic data. The implementation is based on existing DBMSs (askSam and MS Access) and is applied to a database of Austrian German and a database of child speech in four languages. A second article on phonetic databases describes two collections of telephone speech in Swiss French—PolyPhone and PolyVar. The article focuses on data acquisition and the standardization of the annotation of speech data. Furthermore, a verification tool for automatic and semi-automatic processing is introduced. The paper by Edgar Haimerl describes a system for creating dialect atlas maps based on a phonetic database. It can be considered as a front-end for the management and the visualization of phonetic data that were stored in a XBase database. The last article on phonetic data deals with syllable structures. In this project, a database of phonotactic statements was set up for data from more than 200 languages. The software is based on a finite-state automaton using a syllable grammar for each language. The system facilitates search for syllables specific to a given language.

Three articles deal with applications of different linguistic tasks. Andrew Bredenkamp et al. emphasize the management of Russian verb nominalization. The authors describe a database of about 2,000 verbs and their corresponding nominalizations managed by Microsoft's DBMS, MS Access. The paper by Sylviane Granger focuses on the management of a learner corpus for applications in language instruction. The author discusses the use of such a corpus for the identification of common errors by investigating texts by learners and comparing them to texts by native speakers. The work is based on the International Corpus of Learner English (ICLE), which includes about 300,000 words. In the paper by Siobhan Devlin et al., a system for simplifying texts for aphasic readers is introduced. The program is based on the Oxford Psycholinguistic Database and the on-line reference lexicon WordNet. Basically, the program is designed to replace difficult words with more common synonyms to increase the readability of the text. Furthermore, possibilities for syntactic simplifications are discussed in the article.

The last two articles focus on the extension of existing systems. Oliver Christ introduces an approach to the connection of the WordNet thesaurus with the IMS corpus query tool. For this purpose, dynamic attributes have to be declared in the query module, which may be passed to external tools. The article concludes with a detailed discussion of efficiency problems with the use of dynamic attributes. The last article focuses on multilingual data processing in an object-oriented environment. The system used for this purpose is called CELLAR and was developed at the Summer Institute of Linguistics. It is based on object-oriented concepts and can be used to create a "truly multilingual" environment, which means that different languages can be handled in one document at the same time and the system automatically adapts to specific language needs.

In his introduction to the book (18 pages), the editor introduces the contributors, presents the problem, and summarizes the individual papers. The book comprises a selection of papers that were submitted to the conference on linguistic databases in Groningen. The topics discussed in the articles are very task specific and have their origin in very different fields of linguistic research. The papers present informative introductions to the systems they use. The main issue of the book is the description of linguistic data acquisition and organization. The connection to general database design and development is not very strong. The systems, basically, apply quite simple approaches to the storage of linguistic data. The authors propose collections of files with suitable query tools or simple-structured relational databases based on standard software like MS Access or MS FoxPro. However, there are two exceptions to this strategy: the database organization of syllable structures, which includes a grammatical transcription based on finite-state techniques, and the object-oriented system based on the CELLAR environment.

The book does not provide a general view of the topic of linguistic databases and its concepts. However, it should be of interest to people who want to take a closer look at the systems that are introduced in the book.

Jörg Tiedemann is a research assistant at the Department of Linguistics, Uppsala University. His major research interests focus on the automatic processing of parallel text. Currently, he is working on the extraction of lexical translation data from a sentence-linked multilingual corpus, and the organization of the data in a lexical database that is linked to the corpus from which they were extracted. Tiedemann's address is: Uppsala University, Department of Linguistics, Box 527, S-751 20 Uppsala, Sweden; e-mail: joerg@stp.ling.uu.se