Book review

Linguistic Databases. John Nerbonne (ed.)

Esther König, Andreas Mengel
Institute for Natural Language Processing, University of Stuttgart, Azenbergstr. 12, 70174 Stuttgart, Germany, {Esther.Koenig;Andreas.Mengel}@ims.uni-stuttgart.de

BOOK TITLE: Linguistic Databases
EDITOR: John Nerbonne
TOWN OF PUBLICATION: Stanford, Ca.
YEAR OF PUBLICATION: 1998
SERIES: CSLI Lecture Notes Number 77
PUBLISHER: Center for the Study of Language and Information
PRICE IN $ AND/OR £:
PRELIM. PAGES + ARABIC PAGES + INDEXES (COMBINED OR SUBJECT AND/OR AUTHOR'S): 260 pp.
ISBN: 1-57586-092-9

This book is a collection of papers from a conference entitled *Linguistic Databases*, held at the University of Groningen in 1995.

Databases are becoming increasingly important in linguistics and computational linguistics. Databases come in many shapes. Simple text files can be seen as 'text databases', SGML-marked up text is a kind of database, and finally conventional database technology like relational and object-oriented databases can serve to store linguistic utterances and information. Linguistic databases are involved in the conceptual modelling of linguistic data, the convenience and efficiency of retrieval, and availability of technology. This book deals with a variety of specific linguistic modelling tasks: phonological databases, syntactic test suites, lexical semantics, text corpora for language learning, and psycholinguistic data. It consists of twelve individual papers.

Stephan Oepen, Klaus Netter, and Judith Klein describe in *TSNLP - Test Suites for Natural Language Processing* the implementation of a test suite of syntactic phrases. The TSNLP project produced test suites for German, English, and French. Phrases are classified according to rather fine-grained categorial descriptions. In addition, dependency information is given. Furthermore, TSNLP accounts for the fact that syntactic phenomena are interrelated by arranging categories in a network of phenomena. The suites are realized by a relational database scheme. As a consequence, the linguist has to struggle with an SQL-like query language in order to retrieve items from the database. However, this query interface turned out to be well-suited for applying the test suite in a grammar development system. The TSNLP group extended the database scheme with information like flat semantic structures and processing times in order to evaluate their HPSG grammar on a test suite. Converely, errors and inconsistencies in the manually constructed test suites were discovered by checking them against the grammar. Open questions with the TSNLP approach are the non-redundant organisation of families of domain specific test suites, and the integration of test suites into a state-of-the-art grammar development environment which supports e.g. version control.

From Annotated Corpora to Databases: the SgmlQl Language by Jacques Le Maitre, Elisabeth Murisasco, and Monique Rolbert discusses the processing of annotated corpora. The starting point is corpora which are annotated using SGML. For querying these corpora, there are basically two options: converting these corpora into a format that can be read into an SQL database, which is manipulated by SQL commands, or having an SGML compatible data manipulation language. The second approach is described in this contribution. Thus, SgmlQL is an object-oriented extension of SQL for SGML tagged documents, which offers uniformity of encoding and query formulation. In SQL databases, hierarchical relations are more difficult to treat. After an introduction to general concepts of SGML, the operations supported by SgmlQL are motivated and described. The application domain of SgmlQL is hierarchically annotated text corpora. Queries and the results of queries are SGML documents, too. SgmlQL includes operations to restructure results, allows for the retrieval of single elements and the restructuring of hierarchical relations of a document into a new document including insertion, moving, and deletion processes. A processor for SgmlQL is implemented in C and used within the MULTEXT environment. At the end of the chapter, this approach is compared to other query languages. As the availability of annotated corpora grows and SGML and related schemes become standard in this area, the introduction of SGML oriented query languages and the combination and integration of SGML and SQL are highly valuable. The paper gives precise mathematical definitions of SgmlQL. Nevertheless, the definitions could have been accompanied by thorough examples to make clear the full coverage of applications that SgmlQL offers. A good feature of SgmlQL is that the output of queries is written in SGML format, thus resulting corpora can be reused as new resources. As the main focus of application of SgmlQL lie in the field of syntax, most operations apply to structural aspects and SGML elements. SGML attributes and values can only be used as context features, but there are no operations for their manipulation in SgmlQL. If this feature were added, it would certainly enhance the functionality of SgmlQL.

In *Markup of a Test Suite with SGML* by Martin Volk the effective application of SGML to encode linguistic test suites is discussed. The application described is the annotation of syntactic properties of single sentences and various ungrammatical counterparts. As elements and concepts defined in the TEI guidelines are not sufficient, new concepts have to be introduced for this application. For the purpose of effective encoding, three annotation evaluation criteria are defined, namely extensibility, interchangability and redundancy of the material marked up. The goal is to maximize for the first two criteria while minimizing for the third. By providing possible ways of encoding, a stepwise approach to an optimal encoding of the given sentence material is documented. A number of options to encode information of syntactic elements are introduced: elements, attributes, default values of attributes, inheritance mechanisms in SGML and attribute hierarchies. Various solutions are discussed and detailed descriptions of advantages and drawbacks of the approaches are provided. After that, the integration of the formatting, a software system (GTU) used, and query aspects are discussed. At the end of the contribution it is noted that efficient browsing and retrieval software for SGML annotated corpora would be desirable. As the contribution mainly focusses on and exhaustively discusses the solution of the task when using SGML, the exploitability and flexibility of SGML is pointed out very well. However, SGML as such is not questioned. On the contrary, the author provides a lot of solutions to overcome concepts that are not supported by SGML.

An Open Systems Approach for an Acoustic-Phonetic Continuous Speech Database: The S_Tools Database-Management System (STDB-MS) by Werner A. Deutsch, Ralf Vollmann, Anton Noll, and Sylvia Moosmüller addresses technical and procedural aspects of setting up a speech database, accessing it and related problems by describing the work of producing the Austrian German database. After a general introduction about peculiarities of sound processing, choices made for preparing/describing speech signal data including standards on data reduction, sound editing, segmentation, and classification are described. Then interdependencies of technological progress and data to be stored are discussed: As computer facilities are increasingly robust and quick, signal measures (e.g. spectral properties) can be computed in real time by applications and need not be prepared and stored beforehand. Details of the speech material including number and kind of speakers, sentences, proportions of read and spontaneous speech, recording channels and tag information (different linguistic description levels) are given. Possible purposes and applications of the corpus are named:

building lexica, phonetic and morphological research in L1 and L2 language acquisition. Access to the data of the corpus is made possible by a data management system. The novice reader is introduced to the complexity and variety of making a speech database, which needs a lot of linguistic, computational and engineering expertise. However, for the well-versed reader, this contribution offers little new information.

The Reading Database of Syllable Structure by Erik Fudge and Linda Shockey reports on a syllable structure database for the languages of the world. First, the authors motivate the need of the database: For many languages of the world, sound descriptions are available, but most of them only include inventories of sounds (consonants and vowels) and no further phonotactic information or possible tones/accents (phonotactic statements). The authors have collected information on these levels for more than 200 languages. Information of the structures found are documented and represented by phrase-structure-like rules. Thus one can ask for the existence of syllables, words and other complex structures of the relevant languages. Crucial problems of this approach that the authors report, are the necessity of having strict normative rules in order to produce the descriptions from various kinds of sources, of distinguishing between systematic and accidental structural gaps within one language and represent these accordingly, and of identifying and integrating loanwords. The typical application of the database is queries run on the database to test whether certain syllable structures are valid in a particular language or not. Details on the representation, the producers and limitations of the software are given. A problem of the database the authors discuss is that the automata produce more structures than are actually available, so that invalid syllable structures are generated. Even worse is the lack of native language competence when setting up grammars for foreign languages. In general the availability of a collection of the segmental structures of the world is an important resource for phonological research and an ambitious project. Questionable, however, is the positivistic approach to the project: When entering rules for the phonotactical description of the languages under investigation, the result aimed at is already determined. Thus, one has to ask why the entries of the database are not primary data, e.g. phonetic transcriptions of words of the languages described which then can be queried using higher level phonetic expressions.

A Database Application for the Generation of Phonetic Atlas Maps by Edgar Haimerl reports on a project for database driven generation of pronunciation maps. In the project he reports on, questionnaires were used to collect data on Dolomitic Ladinian and its neighbouring dialects. The relevant information of the recorded data was put into a database. The location of the speaker, the orthographic and a phonetic representations, semantic and other linguistic information are stored in the database. A software system (CARD) is used to generate maps, enter, search and correct data, define characters, and vary the layout of the maps, which are printed to PostScript files. Unfortunately, while the author describes some technical details at length, the description of the central aspect of the contribution, i.e. the research purpose, is rather sketchy: No examples are given and thus it is difficult to estimate the value of the software he describes. Some tables of variations of words or sentences would have exemplified the core purpose of the project.

Swiss French PolyPhone and PolyVar: Telephone Speech Databases to Model Inter- and Intra-speaker Variability by Gerhard Chollet, Jean-Luc Cochard, Andrei Constantinescu, Cedric Jabloulet, and Philippe Langlais is a detailed description of a speech data collection project. First, the contribution gives an overview of joint initiatives and consortia in the area of the production and distribution of speech data bases. Next, applications of the collected databases are named: the development, testing, enhancement, and evaluation of speech technology products. The specific purpose of the PolyVar database was to investigate intra-speaker variability, whereas PolyPhone was designed for research in the area of inter-speaker variability. In the subsequent paragraphs, general criteria and data about the PolyPhone and PolyVar databases are presented and discussed: the type and selection criteria of the utterances recorded, the recruitment of speakers, technical information about the recording procedure and speech signal representation, the transcription of the utterances, and the transcription software environment. At the end of the contribution, results of the application i.e. an improvement of a speech recognition system, are presented. Like the contribution by Deutsch et al., this chapter provides an integrated description of research questions, data acquisition, technical problems, and their individual solution.

Investigating Argument Structure: The Russian Nominalization Database by Andrew Bredenkamp, Loiusa Sadler, and Andrew Spencer presents a theory of Russian nominalization, its licensing conditions, and its effects on the predicate argument structures. Several thousand verbs and associated nominal forms have been collected and annotated with predicate-argument structures, aspectual information, and certain morphological information. This collection served to test and to refine Grimshaw's thesis of complex event nominalizations, and to find out relations between certain suffixes and nominalizations. The data was coded in a standard PC-based relational database system, whose graphical user interface facilitated the exploration of the data.

The Use of a Psycholinguistic Database in the Simplification of Text for Aphasic Readers by Siobhan Devlin and John Tait describes the application of the Oxford Psycholinguistic Database and of WordNet for the task of text simplification, in order to make newspaper text comprehensible to people who suffer from language disorders. The Oxford Psycholinguistic Database includes information about the factors which affect lexical comprehension, such as frequency, familiarity, concreteness, age of acquisition. WordNet, from Princeton University, is a thesaurus which defines a variety of lexical relations. For the task of lexical simplification, WordNet delivers synonyms, whose comprehensibility factors are then looked up in the Oxford Psycholinguistic Database. In addition, the paper lists a number of postulates for syntactic simplifications, which are far less trivial to implement. Some preliminary tests with patients indicate an improvement of text comprehension due to the simplified texts.

In *The Computer Learner Corpus: A Testbed for Electronic EFL Tools*, Sylviane Granger argues for the use of both corpora of native English and corpora of learner English in EFL grammar checkers. For a style checker to be useful for a non-native speaker, a mother-tongue specific model of learner competence is required. The paper reports on a project to adapt English style checkers for French-speaking users, on the basis of an adapted version of the International Corpus of Learner English (ICLE). Possible attributes of a learner corpus are mode (written vs. spoken), genre, function, technicality, the learner's mother tongue, and a classification of learners' errors. From such an error-tagged corpus, learner-specific grammar rules can be derived. On the basis of phrase frequencies in native vs. learner language, the learner can be provided with hints not only of errors but also of phrases and collocations which sound more natural. Granger's examples cover faux amis and the use of prepositions, among others. However, the treatment of semantic errors like the generic vs. specific distinction will still require further research. Possibly, the annotation scheme for learners' errors (and the previously described Russian Nominalization database by Bredenkamp et al.) could have profited from the technical contributions of the TSNLP project. Conversely, the builders of test suites for grammar development might take advantage of the knowledge of learners' errors in order to make their grammar test suites more comprehensive.

Oliver Christ's *Linking WordNet to a Corpus Query System* is the description of a flexible corpus query system, which not only supports queries to previously indexed text, but which also allows for the on-line access of corpus-external sources like the WordNet thesaurus. In this way, for example, the contexts for a family of concepts can be searched for simultaneously. Semantically related words can be checked for syntactic properties they may have in common. Christ puts emphasis on the technical aspects of the coupling of a corpus query system with external knowledge sources. The most straightforward approach is a simple Unix pipe, which, however, becomes unacceptably slow when hundreds of thousands of calls to WordNet have to be handled in order to evaluate a query to a corpus of nontrivial size. A client-server set-up has proved to be more reasonable: WordNet is started as a demon ready to process requests. Hence, the paper gives just one example of the use of client-server technology for building a natural language processing system from heterogeneous sources of knowledge.

Multilingual Data Processing in the CELLAR Environment by Gary F. Simons and John V. Thomson is an overview of problems encountered in multilingual dataprocessing. Of particular importance is the handling of scripts for different languages. The authors present the object-oriented CELLAR system, which includes predefined objects and methods for the implementation of multilingual data, such as e.g. a bilingual dictionary or help texts of a computer software. Unfortunately, the paper is missing detailed examples of possible linguistic applications of the CELLAR system. For serious corpus-linguistic uses of the system, the notion of 'multilinguality' should be generalized to 'multilevel' annotation. Furthermore the system should be able to handle the alignment of non-contiguous phrases, i.e. the syntactic divergence in multilingual texts.

In general, the book gives a multifaceted introduction to the technical and theoretical aspects of a variety of ongoing projects in the area that is commonly referred to as *Corpus Linguistics*. Although the stage of the projects in the reports is several years old, one can say that apart from technical improvements in the computer infrastructure, basic theoretical issues are still relevant. As the book includes contributions from different sources, potential addressees are database managers and retrieval experts; speech technologists, phoneticians, and phonologists; theoretical and applied linguists. All of these areas are touched in the book and treated in more or less detail. Obviously, not every member of these groups will benefit from all contributions but it is worthwhile to have a look at the variety offered.