Linguistic Databases

John Nerbonne

March 1, 1996

Contents

Introduction

This is a selection of papers on the use of databases in linguistics. All of the papers were originally presented at a conference entitled "Linguistic Databases", held at the University of Groningen March 23-4, 1995. This introduction reviews the motivation for a special examination of linguistic databases, introduces the papers themselves, and then, briefly, suggests both how the knowledge is useful to working linguists, as well as how databases might evolve to be used for linguistic data more easily.

Motivation

Linguistics is a data-rich study. First, there is a great deal of purely linguistic data. Linguistics sets itself the task of describing and analyzing the structure of language as this is evidenced in billions of speakers, each making hundreds of utterances a day for linguistic lifetimes of several decades. Factors important to the structure of these utterances include thousands of languages, each with tens to hundreds of thousands of words, which in turn may be found in dozens to hundreds of different word forms. The results of linguistic analysis show great variety in the particular coding rules at all levels—sounds, words and phrases. The rules are not individuated clearly enough to allow reliable comparisons, but they seem to number in the thousands in current formulations. The many factors studied in linguistic variation including geography, sex, social and educational status, pathology, and situational "register" (as for telegrams, graffiti, etc.) only add to this embarassment of riches.

Second, various subfields of linguistics have developed experimental methodologies involving a great deal of data which, although not purely linguistic, are used crucially in linguistic theorizing or in applications. Some examples of these would be physcial measurements such as air pressure and pitch, records of social and geographical parameters of variation, the quasi-linguistic ill-formed examples of generative

grammar, or psychological measurements such as reaction time or the movements of eyes in reading.

Third, applications of linguistics need to keep track of further data categories such as successful (and unsuccessful) processings, degree of ambiguity in results, use of particular knowledge sources and heuristics, user reactions, and comparisons to alternative system configurations.

Tasks of Databases

Given this amount of data, it is not surpising that a good number of linguists have investigated software for managing data. Databases have long been standard repositories in phonetics (see Liberman 1997, UCLA Phonetics Laboratory 1996) and psycholinguistics (see MacWhinney 1995) research, but they are finding increasing further use not only in phonology, morphology, syntax, historical linguistics and dialectology but also in areas of applied linguistics such as lexicography and computer-assisted language learning. Normally, they serve as a repositories for large amounts of data, but they are also important for the organization they impose, which serves to ease access for researchers and applications specialists.

The term 'database' refers to collections of electronic records of linguistic data. As a simple example, this might be a file or set of files of sentences. Furthermore, a database records data in a DECLARATIVE form, i.e. independent of the particular procedures needed to interpret, display or modify it. This means that the creators and maintainers of databases have avoided storage forms like the files of word-processing packages, which rely on extensive programs for display, etc.

Similary, something like a parser or generator of sentences is not a database, even if it could be argued to contain the same information. Even if that were so, these forms are by definition abstract and require either a special procedural interpretation, in which case they fail to be declarative, or, in the case of logic grammars, they require sophisticated inference if they are to be used as characterizations of data.[1] The grammars used in parsers and generators are hypotheses about data, and thus serve a completely purpose from databases—even if ideal grammars would characterize the same information. But in fact the information is never the same—the data is not exactly as even the best hypotheses predict.

Furthermore, a database attempts to ensure the CONSISTENCY or IN-

[1] Since the data is always in some sense encoded—there is no direct representation of letters or sounds in electronic circuitry—there are always procedures involved when we look at or otherwise work with the data. But the encoding schemes used are standard and very simple.

TEGRITY of data by eliminating redundant specifications. By eliminating the chance of double specification, we reduce the chance of inconsistent specification. Oepen et al. (below) discuss this requirement in more detail. To give a simple example, if one wished to record the finite verb (of the main clause) in the sentences database, this would NOT best be done by recording it separately (as 'works' in 'Dan works.') since this would introduce the possibility of inconsistency (one might write 'work' or 'wrok'). A preferable solution is to record an index ('2nd'), which is also required to be less than sentence length ($0 < i < length$). Of course, database theory, a vibrant branch of computer science, is only suggested by the remarks above. Further issues arise in making data entry and retrieval easy and fast, in organizing data description (metadata), and in adding logical power (deductive db's) or flexibility (object-oriented db's). Brookshear 1997 contains an excellent introductory chapter on databases—including various data models; and Ullman 1988 and Date 1995 are very popular textbooks, both focusing on relational databases.

The point of the remarks above has been to clarify what's meant by 'database'. If they suggest that the main issues facing the developers and users of linguistic databases today are conceptual, then they are misleading. The most important issues are overwhelmingly practical, as the call for papers suggested. We turn now to a review of the points suggested from the CFP, which called for a forum on the exchange of information and views on the proper use of databases within the various subfields of linguistics. We add notes to each of the points below, suggesting how the conference answered the questions raised. Of course, these are (subjective) assessments.

1. Databases vs. annotated corpora, pros and cons.

   The tendency was less to ask which is correct, amd more to ask how we can have both. LeMaitre et al. and Volk present schemes for viewing annotated corpora as databases. Oepen et al., while adducing advantages of databases, identify the relationship to corpora as a future goal.

2. Needs with respect to acoustic data, string data, temporal data. Existing facilities.

   As the papers by Chollet et al. and Deutsch et al. demonstrate, phonetic databases are already multimedia (and must be), and as Simon and Thompson, Oepen et al. show, string handling is highly desirable in language software and syntactic databases, respectively.

3. Developing (maximally) theory-neutral db schemas for annotation systems.
   The general consensus was that this issue was interesting, but not pressing—as the example of the Penn Treebank Black et al. 1991 showed, even very theory specific annotations are useful. One can use them as comparisons in those parts of analysis that seem right, and they are to some degree translatable into other systems. Even though generality is to be preferred, the work cannot wait until perfect schemes arise.
4. Commercially available systems vs. public domain systems. What's available?
   While some developers were at pains NOT to rely on commercial packages (hoping to keep the threshhold low for users), this was not general. The down-side of this decision is of course the extra development effort (and its duplication). Deutsch et al. and Haimerl discuss packages they use.
5. Uses in grammar checking, replication of results.
   This is a focus in the papers by Volk and Oepen et al.
6. Needs of applications such as lexicography.
   While there was unfortunately no representation from lexicography, Granger and Devlin and Tait present other applications, and applications are discussed in the papers by Oepen et al., Volk, Deutsch et al. and Chollet.
7. Making use of CD-ROM technology.
   Again, little focused discussion even though it figures in the work of Chollet et al. and in the work of the data collection organizations (see following point).
8. Existing professional expertise: Linguistic Data Consortium (LDC),[2] Text Encoding Initiative (TEI) Sperberg-McQueen and Burnard 1994.
   Here we find not only a new organization, the European Language Resources Association (ELRA),[3] but also a very general consensus on the importance of coordinating the overwhelming amount of work being done and yet to be done.

There was very little discussion about a question often heatedly raised in recent meetings between theoreticians and applications specialists: what sorts of data should be compiled and organized—"natural data" versus other. Perhaps this should have been unsurprising since the

[2] See "Introduction to the Linguistic Data Consortium" on the World-Wide Web. There's a pointer from the LDC's homepage at `http://www.cis.upenn.edu/ ldc`

[3] ELRA's director that its address shall be ELRA / ELDA, 87, Avenue D'ITALIE, 75013 PARIS, FRANCE, email: `elra@calvanet.calvacom.fr`

scope of the meeting was much broader than theoretical linguistics, the only subfield in which a RESTRICTION to intuitive data is (sometimes) upheld. Bredenkamp, Sadler and Spencer's contribution demonstrates the value of databases for researchers using intuitive data in generative theory, however, and they, Volk, and Oepen et al. all employ databases for intuitive data. Those references should be enough to correct the misconceptions that the use of databases is tied to particular views of what linguistic data must be or that their use is incompatible with including data about ill-formedness.

On the side of those sceptical about the "natural data" of the data debate, it is worth noting the example of the speech community, who progressed for many years while focusing on the artificial data of speech read from texts with pauses between words. None of this may be interpreted to mean that there was consensus about the status of intuitive vs. natural data, but only that there was no sense of resolving the issue for either side.

A further distinction in the use of databases was deliberately left underspecified in the CFP, and that was the distinction between data and hypotheses (as databases may organize them). Electronic dictionaries and NLP lexicons are in some sense databases, but they are (largely) not databases of data directly, but rather linguistic descriptions of data. Lexicography was mentioned in the CFP in order to solicit contributions from those using database technology to organize linguistic hypotheses. The papers by Chollet et al. and by Deutsch et al. describe systems for the handling of data AND hypotheses in phonetics, and the (unpublished) talks by Gebhardi and Sutcliffe et al. examine the application of database technology to managing lexical information. There was, however, a suggestion that the management of hypotheses (or knowledge bases) puts harsher demands on the need to be able to revise (even though all databases must allow changes).

Papers

In this section we attempt to summarize the individual papers and assess their contributions.

Syntactic Corpora and Databases

Although these papers all arise from the natural language processing field, it is easy to see their relevance to the non-applied, pure science of syntax. There are certainly signs of theoreticians turning to corpora for evidence (Dalrymple et al. 1997).

TSNLP—**Test Suites for Natural Language Processing** by Stephan Oepen, Klaus Netter and Judith Klein presents a detailed

methodology for building a test suite of linguistic examples to be used in evaluating the linguistic coverage of natural language processing systems, and then reports on a large-scale implementation of such test suites for three languages. The authors motivate the development of hand-built test material to complement corpus-based evaluation tools, also suggesting design criteria for test suites. They propose a concrete annotation scheme which tries to remain as neutral as possible with respect to particular linguistic theories. They discuss the broad range of constructions in their test material, and also their portable software for storage and retrieval. In a final section they report on a connecting the test suite to an HPSG grammar and parser.

Markup of a Test Suite with SGML by Martin Volk begins with the Standard Generalized Markup Language (SGML), an international standard which enjoys growing popularity in textual software. Its developers described SGML as "a database language for text", suggesting that it ought to be up to the task of declaratively representing a test suite—an annotated corpus of test sentences, both grammatical and ungrammatical. Volk shows that SGML is indeed suited to this task (discussing five different models) and chooses a model which minimizes redundancy by taking into account the inheritance of syntactic features. The article concludes that SGML is advantageous, but that an optimal inheritance mechanism is still lacking.

From Annotated Corpora to Databases: The SgmlQL Language by Jacques LeMaitre, E. Murisasco and Monique Rolbert describes the SgmlQL query language for SGML documents. This language is an extension of SQL, the standard query language for relational databases. It allows for the retrieval of information from SGML documents and for the modification of markup in SGML documents. The retrieval and modification commands are sensitive to SGML structure and they are made flexible through extensions using regular expressions. The applicability of SgmlQL to various tasks in NLP is discussed.

Phonetic Databases

An Open-Systems Approach for an Acoustic-Phonetic Continuous Speech Database by Werner Deutsch, Ralf Vollmann, Anton Noll and Sylvia Moosmüller was developed to support the design and evaluation of speech recognition. It is a database of acoustic files with a library of database routines to support entry, annotation, editing and update, and retrieval as well as specialized functions for speech such as segmentation, fast-fourier transformations, fundamental and formant frequency extraction, and linear-predictive coding. It is described here as applied to databases of Austrian German (still under construction)

and child language. Its user-interface is implemented as a hypertext system, and, although the work to-date has been confined to the PC platform, the developers are confident of its portability.

Swiss French PolyPhone and PolyVar: Telephone Speech Databases to Model Inter- and Intra-Speaker Variability by Gerard Chollet, J.-L. Cochard, A.Constantinescu, C.Jaboulet and Ph.Langlais reports on the construction of PolyPhone, a database of the telephone speech of 5,000 speakers of Swiss French. The current report focuses on (i) the sampling discipline used, especially how the phonetic material solicited was varied in an effort to minimize collection needed (while guaranteeing a sufficient amount of material in each category); (ii) the transcription of material, including an interface built for this purpose; (iii) labeling procedures, including some automated support. PolyPhone and PolyVar are instances of speech databases being made distributed to the speech research community by the LDC and ELRA.

Applications in Linguistic Theory

Three papers demonstrate applications to questions in linguistic theory.

A Database Application for the Generation of Phonetic Atlas Maps by Edgar Haimerl describes CARD, a system for managing phonetic data to be used to create dialect atlas maps. CARD is a front end to a fairly large (350K entries) PC database (XBase) of data gathered for Salzburg's ALD project (Dolomitic Ladinian). Although CARD handles the basic management of the survey data (input, update, and retrieval), its particular strength lies in its capacity to produce printed maps for arbitrary selections of the data. Because users need to build atlas maps displaying a variety of different features, flexibility is of the utmost importance. CARD achieves this by using the database itself for meta-information. Most application details are stored in database files instead of being hard-coded into the program, and indexes and new databases are created on the fly as the user configures an application. Such techniques help the lexicographer tailor the programme to accommodate specific requirements vis-a-vis selection of data, order of multiple entries, physical appearance of maps, and compound (accented) characters. CARD is expected to be made available when the ALD is published on CD-ROM.

The Reading Database of Syllable Structure by Erik Fudge and Linda Shockey describes a database of syllable structures. The aim of the work is to allow the user to query whether certain syllable types or tokens occur in a language, or indeed to identify languages in the database for which a specified sequence is possible. Thus the database attempts to go beyond the databases such as the UCLA Phonological

Segment Inventory to include the possible combinations of these segments within words, i.e., the phonotactic characteristics of the language. In this paper, the authors report on work in progress. They set up a general database for phonotactic statements for (at the moment) more than 200 languages. Queries supported includes examples such as 'Is [qi] a possible syllable in a certain language or language family?' or 'What percentage of languages allow onsetless syllables?'

Investigating Nominal Argument Structure: the Russian Nominalization Database by Andrew Bredenkamp, Louisa Sadler and Andrew Spencer focuses on the interplay of aspect and argument structure in deverbal nominalization. Following an introduction to aspect and argument linking in syntactic theory and a sketch of Russian verbal morphology, the authors review Grimshaw's distinction between complex event nominalizations on the one hand, and result nominalizations and simple event nominalizations on the other, in which only complex event nominalizations are claimed to have argument structure. Russian is selected to test the hypothesis because some tests rely on aspect, which is overtly marked in Russian. To organize the relevant data, the authors have organized it into a database which contains 2,000 Russian verbs and their corresponding nominalizations. In the last section of the paper, the authors discuss a first result: while underived imperfective verbs and simplex and derived perfective verbs give rise to both complex event nominalizations and result and simple event nominalizations, derived imperfective verbs give rise to complex event nominalizations only. This had not been noticed in purely theoretical work.

Applications

The Computer Learner Corpus: A Testbed for Electronic EFL Tools, an invited paper by Sylviane Granger, reports on progress in language instruction based on the International Corpus of Learner English, a corpus collected from the (of course imperfect) language of learners. Earlier improvements in language instruction had resulted from the careful use of native-speaker corpora, e.g., more attention to frequent words in introductory texts. The present report demonstrates the use of corpora in identifying common errors and in assessing the worth of computerized tools, especially grammar checkers, for language learners—where there seems to be a need for more specialized software. The paper concludes with a plea for the adaptation of existing software to this application area.

The Use of a Psychlinguistic Database in the Simplification of Text for Aphasic Readers by Siobhan Devlin and J.Tait reports on the immplementation of a program to simplify text which makes use

of both the Oxford Psycholinguistic Database as well as WordNet, a dictionary database structured semantically. The program is designed to replace difficult words by more common synonyms. The Oxford Psycholinguistic Database contains information on frequency, concreteness and other measures which contribute to the difficulty of word, and WordNet contains information on synonyms. The program is intended for use in testing what factors improve the ease with which aphasics can read.

Extending Basic Technologies

Two papers provide suggest some of the directions in which linguistic databases of the future may lie.

Linking WordNet to a Corpus Query System by Oliver Christ addresses the general problem of increasing the utility of text corpora. The strategy proposed is to consult knowledge bases external to the corpus to allow more precise queries. The author notes that the standard approach to annotating corpora is to add static attributes to the wordforms, and that his proposal differs in that it defines "dynamic attributes" which query external sources to compute more annotations on-the-fly. The approach is exemplified by showing how a dynamic link to the WordNet thesaurus makes it possible to constrain searches on a text corpus by semantic information computed from WordNet for the wordforms.

Multilingual Data Processing in the CELLAR Environment, an invited paper by Gary F. Simons and John V. Thomson addresses a problem which researchers have wished would go away more than they have tried to solve: the atavistic "monolinguality" of the computer world. CELLAR is a "comprehensive environment for linguistic, literary and anthropological research", including innovative ideas on what is needed in linguistic databases, but, at the urging of the program committee, the authors agreed to focus here on the multilingual issues in system design. In spite of encouraging recent improvements (mostly in displaying foreign texts), we still have trouble with inputing from the keyboard, spell-checking, sorting (e.g., in alphabetic order), searching, embedding within other languages, or cross-referencing to elements (alignment, as used in glossing or parallel texts). A fundamental requirement within CELLAR is that each string data element must be marked for the language it is written in. Multilingualism extends beyond data: the user interface is likewise multilingual, so that menus and button labels, error messages etc. are available in several languages.

Going on

How can the information here be useful to working linguists? Most directly, of course, if the papers here provide some ideas on how to get data and on how to organize what data you have so that you and others can get the most out of it. The organizations to turn to have been mentioned above: LDC, TEI and ELRA.

More databases in general:

`http://www.cis.ohio-state.edu/hypertext/faq/usenet/databases/free-databases/faq.html` is a catalogue of free database systems.

The Conference

There were several presentations at the conference which do not appear below. The novelty of the topic made it was impossible to be certain beforehand whether there would be work of sufficient interest and quality to warrant publication. The original CFP therefore explicitly refrained from promising publication, but the meeting demonstrated that this would be valuable. Unfortunately, some authors had commitments elsewhere, etc. which prevented their contributions from being included here. Undoubtedly the collection would be even more valuable if more of the conference contributions below could have been included:

Susan Armstrong (ISSCO, Geneva) and Henry Thompson (Edinburgh) A Presentation of MLCC: Multilingual Corpora for Cooperation

Dietmar Zaefferer (Munich) Options for a Cross-Linguistic Reference Grammar Database

Masahito Watanabe (Meikai, Yokohama) A Better Language Database for Language Teaching

Gunter Gebhardi (Berlin) Aspects of Lexicon Maintenance in Computational Linguistics

Richard Sutcliffe et al. (Limerick) From SIFT Lexical Knowledge Base to SIFT Lexical Data Base: Creating a Repository for Lexicological Research and Development

Pavel A. Skrelin (St. Petersburg) The Acoustic Database for Studying Language Changes

Kamel Bensaber, Jean Serignat and Pascal Perrier (ICP, Grenoble) BD_ART: Multimedia Articulatory Database

Lou Boves and Els den Os (SPEX, Leidschendam) Linguistic Research using Large Speech Corpora

In particular, the conference benefitted from the presentations of Jan Aarts and Mark Liberman, who were invited to speak on their decades of accomplishments in this area:

Jan Aarts Prof. of English, Nijmegen, leader of the TOSCA, and Linguistic Databases projects: "Annotation of Corpora: General Issues and the Nijmegen Experience"

Mark Liberman Prof. of Linguistics & Computer Science, University of Pennsylvania and Director, Linguistic Data Consortium "Electronic Publication of Linguistic Data"

Finally, there were demonstrations by Judith Klein (TSNLP), Gunter Gebhardi (LeX4), Edgar Haimerl (ALD), Werner Deutsch (S_TOOLS) and Hans van Halteren (LDB). The TSNLP, ALD and S_TOOLS work is reported in the papers (coauthored) by the demonstrators.

Acknowledgments

The editor wishes to thank the program committee, Tjeerd de Graaf, Tette Hofstra and Herman Wekker for scientific and organizational advice; Duco Dokter and Edwin Kuipers, who handled local arrangements; and Sietze Looyenga, who managed finances.

The conference was supported financially by grants from the Dutch National Science Foundation (NWO), Royal Dutch Academy of Science (KNAW), the Groningen Center for Language and Cognition (CLCG), and the Behavioral and Cognitive Neurosciences Graduate School, Groningen (BCN). Our thanks to all of these organizations. The editor further wishes to thank the Japanese Ministry of Post and Telecommunciations, who sponsored a visit during which this introduction was written.

John Nerbonne
Groningen, 3/96

Bibliography

Black, Ezra, Stephen Abney, Dan Flickinger, C. Gdaniec, Ralph Grishman, Phil Harrison, Don Hindle, Robert Ingria, Fred Jellinek, Judith Klavans, Mark Liberman, Mitch Marcus, Salim Roukos, Beatrice Santorini, and Tomas Strzalkowski. 1991. A Procedure for Quantitatively Comparing the Syntactic Coverage of English Grammars. In *Proc. of the Feb. 1991 Speech and Natural Language Workshop*. San Mateo: Morgan Kaufmann.

Brookshear, Glenn. 1997. *Computer Science: An Overview*. Reading, Massachusetts: Addison-Wesley. 5th ed.

Dalrymple, Mary, Makoto Kanazawa, Yookyung Kim, Sam Mchombo, and Stanley Peters. 1997. Reciprocal Expressions and the Concept of Reciprocity. *Linguistics and Philosophy*. accepted for publication.

Date, C. J. 1995. *An Introduction to Database Systems*. Reading, Massachusetts: Addison-Wesley. 6 edition.

Liberman, Mark. 1997. Introduction to the Linguistic Data Consortium. available via the LDC site: `http://www.ldc.upenn.edu/`.

MacWhinney, Brian. 1995. *The CHILDES Project: Tools for Analyzing Talk*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Sperberg-McQueen, C.Michael, and Lou Burnard. 1994. *Guidelines for Electronic Text Encoding and Interchange*. Oxford and Chicago: Text Encoding Initiative. Version 3 (first publicly available version).

UCLA Phonetics Laboratory. 1996. The Sounds of the World's Languages. See `http://www.humnet.ucla.edu/humnet/linguistics/faciliti/`.

Ullman, Jeffrey D. 1988. *Principles of Database and Knowledge-Base Systems*. Rockville, Maryland: Computer Science Press.