Electronic Databases for Linguistic and Language Research

Beatrice T. Oshika and Sylvia C. Krausse

Abstract

Machine-readable language data have increased enormously because of computer-based publishing, online record-keeping, online generation of business and government documents, and electronic mail, to mention just a few sources. This article reviews electronic language data oriented specifically toward linguistic and language research, including dictionaries, other text material, speech databases, bulletin boards, and news groups available over common networks.

Introduction

The availability of machine-readable language data is growing rapidly because of computer-based publishing, record keeping, and documentation. The enormous amount of information represented in natural language serves as a rich resource for research about natural language. In recent years, the accessibility of machine-readable language data has begun to shift basic approaches in language analysis toward corpus-based research, describing linguistic patterns—especially lexical distributions—as they are reflected in actual data. These large databases have generated increased interest in lexicography and also in statistical approaches to language analysis, a trend that is documented in Aarts and Meijs (1990), Church (1991), and Garside, et al. (1987).

This article describes some of the major sources of such data, focusing on large databases that are commonly available, widely used,
and of broad interest to linguists and language researchers. The emphasis is on sources that could serve general needs of librarians, scholars, and students. Not included in this survey are instructional materials such as specific computer-assisted language learning programs and editing and formatting programs to handle nonroman scripts and specialized fonts. Also omitted here is mention of standard reference databases already well known and easily available through DIALOG or BRS such as *Linguistics and Language Behavior Abstracts* (LLBA) (see DeMiller [1991] for a concise discussion on the major online and CD-ROM databases).

Four main types of electronic databases are included:

1. machine-readable versions of established dictionaries;
2. running text (e.g., newspapers, books);
3. speech data with phonetic/orthographic transcriptions; and
4. newsgroups and bulletin boards available over networks

**Dictionaries**

Virtually all modern dictionaries are produced electronically and are available in both print and machine-readable form, but that situation is very recent. Just over twenty years ago, the first machine-readable version of a dictionary was *Webster's Seventh Collegiate Dictionary* which had been generated by key-punch entry by John Olney (1968) as part of his philosophical research on word senses. A review of the development of electronic lexical databases since that time can be found in Evens (1988, pp. 9-16).

Essentially all American dictionaries are now available on CD-ROM or magnetic tape or through online access with supporting search software. A recent article in *CD-ROM Librarian* (Urrows & Urrows, 1991) provides an excellent review of electronic dictionaries and other language aids, a few of which are described here.

*Webster's Ninth New Collegiate Dictionary* is available on CD-ROM compatible with the Macintosh Plus, Macintosh II, and Macintosh SE and includes optional audio output with recordings of all words.

The *American Heritage Electronic Dictionary* is available in a licensing arrangement from Houghton Mifflin in three versions (comprehensive, concise, and compact), along with *Roget's Thesaurus*, and language and grammar tools. Houghton Mifflin also offers online access to a computerized version of the *Random House Dictionary*.

The *Longman Dictionary of Contemporary English* (LDOCE) is notable because its grammar coding system is the "most comprehensive description of grammatical properties of words to be
found in any published dictionary in machine readable form” (Boguraev & Briscoe, 1987, p. 203). The LDOCE has 60,000 entries which are labeled with relatively rich grammatical subcategorizations and has considerable information on phrasal verbs, noun compounds, and idioms generally not found in dictionaries.

The CD-ROM version of the Oxford English Dictionary (OED) represents the twelve volumes of the 1933 edition. It comes with eight indexes and support software for the IBM PC/XT/AT/PS2. The developers intend to maintain the dictionary indefinitely and make it available in both print and machine-readable form (Simpson, 1985).

Languages of the World is a composite multilingual dictionary on CD-ROM containing material from eighteen bilingual dictionaries covering twelve languages plus the NTC Comprehensive Dictionary of American Idioms. The languages included are Chinese, Danish, Dutch, English, Finnish, French, German, Italian, Japanese, Norwegian, Spanish, and Swedish. The software allows searching and selecting on any combination of languages and supports linkage with common word processing programs so that dictionary entries can be inserted into a document.

**Text Databases**

Corpora representing running text are an important source of language data. Such databases are generally derived from printed sources of running narratives and may include newspapers, books, and journals. In addition to the text itself, these databases may include concordances and grammatical tags (e.g., noun, adjective) for each lexical item.

One of the most significant American lexicographic efforts was assembled at Brown University in the 1960s. Kucera and Francis (1967) collected 2,000 words each from a variety of newspapers, books, and periodicals to construct a 1 million word machine-readable corpus of modern American English, the largest of its type at the time. Over several years, intensive work by staff and graduate students resulted in concordances for each word and grammatical tags. Frequency lists of the Brown corpus are available in many forms—e.g., it is included as a part of the MRC Psycholinguistic Database available at handling cost from the Oxford Text Archive (e-mail: <archive@vax.ox.ac.uk>).

The Brown corpus has been the basis for many language processing studies in the past twenty-five years and continues to play an important role in lexical research. However, a million word database is now considered quite small, especially as a foundation for significant statistical analyses.

The British equivalent of the Brown corpus is the Lancaster/Oslo-Bergen (LOB) corpus described in a two-volume set analyzing
tag and word frequencies and combinations (Johansson & Hofland, 1989). Sampling methods and text categories are similar to those of the Brown project, and the database continues to play a similarly significant role in studies of modern British English. The corpus is distributed by the International Computer Archive of Modern English (ICAME) based in Bergen, Norway (contact Knut Hofland, e-mail: <fafkh@earn.nobergen>).

A larger database for British English (and including German and Dutch) is the CELEX lexical database being developed at the Centre for Lexical Information within the University of Nijmegen (e-mail: <celex@celex.kun.nl>). Based on the 18.7 million words of the Birmingham corpus, CELEX is distributing word frequency lists for 100,000 British English headwords and their inflections. The database represents extracts from mainly British English texts published between the 1950s and the early 1980s.

Also representative of current larger databases is the British COBUILD (Collins Birmingham University International Database) database being developed by the University of Birmingham, Birmingham, England, in cooperation with Collins Publishers (Sinclair, 1987b). The total corpus is 20 million words collected from books, magazines, newspapers, and journals plus transcripts of speeches from radio interviews. All materials were written after 1960 so the database is representative of modern English. Most of the sources are British, but specific efforts were made to include data from Australian, South African, Indian, and American English.

Concordances are available for the entire COBUILD database along with the source so that subsets of data representing dialect or style can be identified. This project required development of a large number of lexicography tools for database entry and access, and a learners dictionary has been published (Simpson, 1987a).

An even larger database is being developed through the Data Collection Initiative (DCI) established by the Association for Computational Linguistics in early 1989 (Liberman, 1989). More than 300 million words of text have been acquired so far, with an initial release of a 30 million word sample. Materials include the 1979 edition of the Collins English Dictionary; articles from the Wall Street Journal; parallel bilingual French-English translations of Canadian parliamentary proceedings (the so-called Hansard database); scientific abstracts; about a dozen novels and plays; the Bible; economic reports from a Swiss bank; the Swiss Civil Code in French, German, and Italian, and at least one speech sample of a National Public Radio news report with orthographic, phonetic, and prosodic transcriptions.

Coding and annotation for this project is being coordinated through the Text Encoding Initiative (TEI) jointly sponsored by the
ACL, the Association for Computers and the Humanities, and the Association for Literary and Linguistic Computing (ACL, 1990). Funding for the TEI comes from the National Endowment for the Humanities and the European Economic Community, and the goal is to define standards for coding and tagging a wide range of texts to encourage sharing of data and cooperative research efforts. One of the main issues arising in electronic lexicography is the need for such standards in the face of the diverse array of formats and notational conventions in which data has been coded.

Over fifty scholars from Europe, the Middle East, and North America are participating in the TEI efforts to define sets of tags for marking features of texts and to code the tag sets within the framework of the Standard Generalized Markup Language (SGML). This will allow marking of physical features of text, such as character sets and page layout, and also linguistic features, such as metrical structure or syntactic constructions.

Information for tape requests on the ACL/DCI and TEI initiatives is available from Donald Walker at the Association for Computational Linguistics (e-mail address: <walker@flash.bellcore.com> or <dci@flash.bellcore.com>).

**Speech Databases**

Speech recognition and synthesis efforts in the 1980s plus advances in storage technology have led to establishment of shared speech databases that can be used for algorithm development and system evaluation and testing. As speech database collection is both labor and computationally intensive, the accessibility to shared data can considerably accelerate research progress. Such databases also provide a common reference for evaluation across systems and techniques.

The issue of standardization is even more acute in speech data collection because of the significant effects of microphone, sampling frequency, and recording environment on the data and on the algorithms with which they will be used. In addition, standard alphabets and notational conventions for orthographic and phonetic transcriptions and time-aligned labeled acoustic data are also needed. Speech data differ considerably from text data, with pauses, interjections, and other hesitation phenomena which affect the realization of the spoken language. These standards issues are reviewed in Fourcin et al. (1989).

The DARPA-supported spoken language systems research has resulted in several speech databases, including the large TIMIT database developed by Texas Instruments and MIT. The database represents ten utterances elicited from 630 male and female speakers...
for a total of 6,300 sentences. The utterances are phonetically transcribed using semi-automatic labeling tools and hand correction. An inventory of sixty standard International Phonetic Association symbols is available for transcription. This database is being used in ongoing speech recognition research and is available from the National Institutes for Science and Technology (formerly the National Bureau of Standards).

The JEIDA Japanese language database includes data from 150 people—75 male and 75 female—speaking monosyllables, discrete words, and connected digits, for a total of 323 utterances, repeated four times by each speaker, reading from word lists.

A similar database for French is GRECO, representing speech from thirty-two speakers, including words; phrases; hundreds of figures, numbers, and number sequences; letters and spelled words; and several hundred consonant-vowel utterances.

The European ESPRIT project collected data from four speakers in each of five languages (Danish, Dutch, English, French, and Italian) for a total of twenty speakers. Utterances included single digits and digit sequences. The Centre for Speech Technology Research at Edinburgh collected data from sixty speakers, male and female, who read a pre-written speech plus a prose reading passage. A NATO project recorded digits and digit sequences from nineteen speakers of British English, American English, French, German, and Dutch, for a total of 1,400 isolated digits and 25,000 digit sequences. It is clear that the availability of speech databases is not as advanced as that of text databases because of the logistics and resources associated with speech recording under controlled conditions.

### Network Resources and Other Sources

The availability of international electronic networks, such as Bitnet and Internet, has fundamentally changed communications among scholars and students in linguistics and languages. Direct contact with colleagues, immediate answers, and global broadcasting of queries and responses all serve to support interchange of data and research results.

In addition to newsgroups on various topics available on all networks to all users, there are focused electronic bulletin boards oriented toward the language scholar. One of these is LN, an international distribution list for computational linguists. The list has several hundred members worldwide and serves as a forum to exchange information, including calls for papers, conference announcements, project descriptions, and requests for data or bibliographic references. Sponsors are the Association for Computational Linguistics and the Association for Computers and the
Humanities. Members subscribe (at no cost) and receive all postings. Moderators are Jean Veronis (e-mail: <veronis@vassar.bitnet>) and Peter Zweigenbaum (e-mail: <zweig@frsim5l.bitnet>). Notices are in French or English and often provide very useful contacts about research in French language and linguistic studies in particular and international efforts in general.

To address the issue of tools for handling large amounts of text data, a research consortium has been established at the Computing Research Laboratory of New Mexico State University. The Consortium for Lexical Research is sponsored by the Association for Computational Linguistics and funded by the Defense Advanced Research Projects Agency (DARPA). Members of the consortium will pay an annual fee and will share lexical data and tools used for research on natural language dictionaries and lexicons and communicate results of the research. Participants contribute resources to a repository and withdraw other resources that support their research. A major activity of the consortium will be to negotiate agreements with providers of tools for the mutual advantage of both suppliers and researchers. Information on the consortium is available by e-mail from: <lexical@nmsu.edu>.

Sample current materials include Roget's 1911 Thesaurus, the Wordsmyth 47,000 entry dictionary, several major natural language processing software tools from Bolt, Beranek and Newman, and sample data collected as part of the Text Encoding Initiative.

Another source of information about language processing tools and other natural language software (as distinguished from lexical corpora) is the Natural Language Software Registry (e-mail: <registry@tira.uchicago.edu>). To make evaluation and exchange of natural language software easier, the University of Chicago's Center for Information and Language Studies has begun to catalog both research and commercial products. Developers are asked to submit descriptions, including type of system (e.g., commercial or research), application (e.g., machine translation, database query system, parser), system components, hardware and software platform and other system specifications, availability and distribution, and restrictions on use.

The Bibliographic Database (BDB), focused specifically on computer processing of language data, is available from the journal Paralingua based in Montreal. As of August 1989, there were approximately 38,000 international citations in the BDB representing 21,000 different authors. The BDB covers a wide range of topics, including character recognition, document processing, information retrieval from text, and literary computing.
CONCLUSION

Computer technology is contributing significantly to development of language databases to support linguistics and language research. Because of the increasing size of the databases, more data sharing, and better communication among researchers and system developers, several major issues have emerged.

1. The availability of data is actually changing the methodology that language researchers use. Techniques that were intractable a decade ago because of constraints on computing resources and on data can now be employed. Statistical inference (or bottom up or data driven) techniques, such as Hidden Markov Models and n-gram statistics, can be tried in computational linguistic applications such as machine translation which traditionally has relied on top down methods. Recent years have seen an explosion of interest in lexicography, a direct result of computer-based publishing.

2. Sharing of data and sharing of software tools requires a framework of standardization in collection, transcription, and maintenance of data, issues being addressed by the Text Encoding Initiative. Not only are the text strings themselves important but also spatial information such as placement on the page and order of characters (not left to right in all languages, not horizontal in all languages).

3. Access to common networks has made the research community closer, more communicative, and more likely to work in a collaborative way. The contribution of electronic mail, bulletin boards, and news groups cannot be overestimated. The print medium or the telephone is no longer the way to keep current with research peers.

4. The nature of professional publication is also changing as the value of high quality databases is acknowledged. Scholars will count database development and dissemination on a par with journal articles, and professional achievement will be measured based on sharing of data as well as research results.

REFERENCES


