Chapter 12

Language Resources

12.1 Overview

John J. Godfrey[a] **& Antonio Zampolli**[b]

[a] Texas Instruments Speech Research, Dallas, Texas, USA
[b] Istituto di Linguistica Computazionale, CNR, Pisa, Italy

The term *linguistic resources* refers to (usually large) sets of language data and descriptions in machine readable form, to be used in building, improving, or evaluating natural language (NL) and speech algorithms or systems. Examples of linguistic resources are written and spoken corpora, lexical databases, grammars, and terminologies, although the term may be extended to include basic software tools for the preparation, collection, management, or use of other resources. This chapter deals mainly with corpora, lexicons, and terminologies.

An increasing awareness of the potential economic and social impact of natural language and speech systems has attracted attention, and some support, from national and international funding authorities. Their interest, naturally, is in technology and systems that work, that make economic sense, and that deal with real language uses (whether scientifically *interesting* or not).

This interest has been reinforced by the success promised in meeting such goals, by systems based on statistical modeling techniques such as hidden Markov models (HMM) and neural networks (NN), which learn by example, typically from very large data sets organized in terms of many variables with many possible values. A key technical factor in the demand for lexicons and corpora, in fact, is the enormous appetites of these techniques for structured data. Both in speech and in natural language, the relatively common occurrence of relatively uncommon events (triphones, vocabulary items), and the disruptive effect of even minor unmodeled events (channel or microphone differences, new vocabulary items, etc.) means that, to provide enough examples for statistical methods to work, the corpora must be numerous (at the very least one per domain or application), often massive, and consequently expensive.

The fact that we still lack adequate linguistic resources for the majority of our languages can be attributed to:

- The tendency, predominant in the '70s and the first half of the '80s, to test linguistic hypotheses with small amounts of (allegedly) critical data, rather than to study extensively the variety of linguistic phenomena occurring in communicative contexts;
- The high cost of creating linguistic resources.

These high costs require broadly-based cooperative efforts of companies, research institutions and sponsors, so as to avoid duplications and to widely share the burden involved. This obviously requires that linguistic resources not be restricted to one specific system, but that they be *reused*—by many users (*shareable* or *public* resources), or for more than one purpose (*multifunctional* resources). There are many examples of the former, such as the TIMIT corpus, TI-DIGITS, Treebank, the Celex Lexical Database, the Italian machine dictionary, and a few of the latter, such as SWITCHBOARD (used for speaker identification, topic detection, speech recognition, acoustic phonetic studies), the GENELEX dictionaries and the MLCC corpus.

A controversial problem, especially with natural language materials, is whether, in order to be reusable and multifunctional, linguistic resources must also be *theory-neutral*: the requirements for linguistic information of a given natural language or speech system may depend not only on the intended applications, but also on the specific linguistic theories on which the system's linguistic components are explicitly or implicitly based.

At the scientific and technical level, the solution is to attempt a consensus among different theoretical perspectives and systems design approaches. Where successful, this permits the adoption of common specifications and de facto standards in creating linguistic resources and ensures their harmonization at the international and multilingual level. The Text Encoding Initiative, jointly sponsored by ACH (Association for Computing in the Humanities), ALLC (Association of Literary and Linguistic Computing), and ACL (Association for Computational Linguistics), has produced a set of guidelines for encoding texts. The project LRE-EAGLES (Expert Advisory Group on Linguistic Engineering Standards), recently launched by the CEC DGXIII, is pooling together the European efforts of both academic and industrial actors towards the creation of de facto consensual standards for corpora, lexicons, speech data, and for evaluation and formalisms.

At the organizational level we can recognize, with regard to the present state of the art, the need for three major action lines:

(a) to promote the reuse of existing (partial) linguistic resources. This can imply various tasks, from reformatting or converting existing linguistic resources to common standards, to augmenting them to comply with common minimal specifications, to establishing appropriate agreements for putting some resources in the public domain;

(b) to promote the development of new linguistic resources for those languages and domains where they do not exist yet, or only exist in a prototype stage, or exist but cannot be made available to the interested users; and

(c) to create cooperative infrastructure to collect, maintain, and disseminate linguistic resources on behalf of the research and development community.

The most appropriate way to organize these activities is still under discussion in various countries.

In Europe, the CEC DG-XIII LRE-RELATOR project, begun in 1995, aims at creating an experimental organization for the (c) tasks. The LE-MLAP (Language Engineering Multilingual Action Plan) has launched projects for activities of type (a) and (b) in the field of written and spoken corpora, lexicons, and terminology.

In Japan, plans for a central organization for speech and text databases have been under discussion. The EDR (Electronic Dictionary Research) Institute is, a the time of the writing of this volume, about to conclude the creation of large monolingual Japanese and English lexicons, together with bilingual links, a large *concept* dictionary and associated text corpora.

The approach taken in the U.S. was to create the Linguistic Data Consortium (LDC); although started with a government grant, it depends on membership dues and data collection contracts for its continued operations. LDC's principal mission is exactly (c) above, but in fulfilling the needs of its worldwide membership it addresses (a) and (b) as well. In its first three years it has released over 275 CD-ROMs of data for public use. Examples of its activities include:

- Publication of existing corpora previously available only to government contractors;

- Collection of speech and text data in languages of interest to members (English, Mandarin, Japanese, Spanish, French, and others);

- Creation of Common Lexical Databases for American English and other languages, with free commercial licenses for members;
- Acting as a clearinghouse for intellectual property rights to existing linguistic resources;
- Campaigning for the release of government-owned resources to researchers.

The need for ensuring international cooperation in the creation and dissemination of linguistic resources seems to us a direct consequence of their infrastructural role, precompetitive nature, and multilingual dimension. The CEC is taking a leading role for the coordination, among the EU countries and EU languages. COCOSDA (for speech) and LIRIC (for NL) are spontaneous initiatives of the R&D international community which aim at ensuring world-wide coordination. Inside the framework of EAGLES and RELATOR, the possibility of defining a common policy for cooperation between the major sponsoring agencies (CEC, NSF, ARPA, MITI) is being explored.

12.2 Written Language Corpora

Eva Ejerhed[a] & Ken Church[b]

[a] University of Umea, Sweden
[b] AT&T Bell Labs, Murray Hill, New Jersey, USA

12.2.1 Review of the State of the Art in Written Language Corpora

Written Language Corpora, collections of text in electronic form, are being collected for research and commercial applications in natural language processing (NLP). Written Language Corpora have been used to improve spelling correctors, hyphenation routines and grammar checkers, which are being integrated into commercial word-processing packages. Lexicographers have used corpora to study word use and to associate uses with meanings. Statistical methods have been used to find interesting associations among words (collocations). Language teachers are now using on-line corpora in the classroom to help learners distinguish central and typical uses of words from mannered, poetic, and erroneous uses. Terminologists are using corpora to build glossaries to assure consistent and correct translations of difficult terms such as *dialog box*, which is translated as *finestra 'window'* in Italian and as *boite 'box'* in French. Eurolang is currently integrating glossary tools, translation memories of recurrent expressions, and more traditional machine translation systems into Microsoft's Word-for-Windows and other popular word-processing applications. The general belief is that there is a significant commercial market for multilingual text processing software, especially in a multilingual setting such as the European Community. Researchers in Information Retrieval and Computational Linguistics are using corpora to evaluate the performance of their systems. Numerous examples can be found in the proceedings of recent conferences like the Third Message Understanding Conference (DARPA, 1991b), and the Speech and Natural Language Workshops sponsored by the Defense Advanced Research Projects Agency (DARPA) (DARPA, 1992a; ARPA, 1993a; ARPA, 1994).

Written language corpora provide a spectrum of resources for language processing, ranging from the *raw material* of the corpora themselves to *finished components* like computational grammars and lexicons. Between these two extremes are intermediate resources like annotated corpora (also called tagged corpora in which words are tagged with part of speech tags and other information), tree banks (in which sentences are analyzed syntactically), part-of-speech taggers, partial parsers of various kinds, lexical materials such as specialized word lists and listings of the constructional properties of verbs.

The corpus-based approach has produced significant improvements in part-of-speech tagging. Francis and Kucera (1982) enabled research in the U.S. by tagging the Brown Corpus and making it available to the research community. Similar efforts were underway within the International Computer Archive of Modern English (ICAME) community in the UK and Scandinavia around the same time. A number of researchers developed and tested the statistical *n-gram* methods that ultimately became the method of choice. These methods used corpora to train parameters and evaluate performance. The results were replicated in a number of different laboratories. Advocates of alternative methods were challenged to match the improvements in performance that had been achieved by *n-gram* methods. Many did, often by using corpus-based empirical approaches to develop and test their solutions, if not to train the parameters explicitly. More and more data collection efforts were initiated as the community began to appreciate the value of the tagged Brown Corpus.

Of course, corpus analysis is not new. There has been a long empirical tradition within descriptive linguistics. Linguists have been counting words and studying concordances for hundreds of years. There have been corpora, libraries and archives for as long as there has been written language. Text has been stored in electronic form for as long as there have been computers. Many of the analysis techniques are based on Information Theory, which predates computers.

So why so much interest, and why now? The role of computers in society has changed radically in recent years. We used to be embarrassed that we were using a million dollar computer to emulate an ordinary typewriter. Computers were so expensive that applications were supposed to target exclusive and unusual needs. Users were often expected to write their own programs. It was hard to imagine a computer without a compiler. Apple Computer Inc. was one of the first to realize that computers were becoming so cheap that users could no longer afford to customize their own special-purpose applications. Apple took a radical step and began to sell a computer without a compiler or a development environment, abandoning the traditional user-base and targeting the general public by developing user-friendly human-machine interfaces that anyone could use. The emphasis moved to so-called *killer* applications like word-processing that everyone just had to have. Many PCs now have email, fax and a modem. The emphasis on human-machine interfaces is now giving way to the information *super-highway* cliche. Computers are rapidly becoming a vehicle for communicating with other people, not very different from a telephone.

> "Phones marry computers: new killer applications arrive."
> – cover of *Byte* magazine, July 1994

Now that so many people are using computers to communicate with one another, vast quantities of text are becoming available in electronic form, ranging from published documents (e.g., electronic dictionaries, encyclopedias, libraries and archives for information retrieval services), to private databases (e.g., marketing information, legal records, medical histories), to personal email and faxes. Just ten years ago, the one-million word Brown Corpus (Francis & Kucera, 1982) was considered large. Today, many laboratories have hundreds of millions or even billions of words. These collections, are becoming widely available, thanks to data collection efforts such as the following: Association for Computational Linguistics' Data Collection Initiative (ACL/DCI), the Linguistic Data Consortium (LDC),[†] the Consortium for Lexical Research (CLR),[†] the Japanese Electronic Dictionary Research (EDR), the European Corpus Initiative (ECI),[†] International Computer Archive of Modern English (ICAME),[†] the British National Corpus (BNC),[†] the French corpus Frantext of Institut National de la Langue Francaise (INaLF-CNRS),[†] the German Institut für deutsche Sprache (IDS),[†] the Dutch Instituut voor Nederlandse Lexicologie (INL),[†] the Danish Dansk Korpus (DK),[†] the Italian Istituto di Linguistica Computazionale (ILC-CNR),[†] the Spanish Reference Corpus Project[†] of Sociedad Estatal del V Centenario, Norwegian corpora of Norsk Tekstarkiv,[†] the Swedish Stockholm-Umea Corpus (SUC)[†] and corpora at Sprakdata, and Finnish corpora of the University of Helsinki[†] Language Corpus Server. This list does not claim to be an exhaustive listing of data collections or data collection efforts, but an illustration of their breadth. Data collections exist for many languages in addition to these, and new data collection efforts are being initiated. There are also standardization efforts for the encoding and exchange of corpora such as the Text Encoding Initiative (TEI).[†]

12.2.2 Identification of Significant Gaps in Knowledge and/or Limitations of Current Technology

The renaissance of interest in corpus-based statistical methods has rekindled old controversies—rationalist vs. empiricist philosophies, theory-driven vs. data-driven methodologies, symbolic vs. statistical techniques. The field will ultimately adopt an inclusive strategy that combines the strengths of as many of these positions as possible.

In the long term, the field is expected to produce significant scientific insights into language. These insights would hopefully be accompanied by corresponding accomplishments in language engineering: better parsers, information retrieval and extraction engines, word processing interfaces with robust grammar/style checking, etc. Parsing technology is currently too fragile, especially on unrestricted text. Text extraction systems ought to determine who did what to whom, but it can be difficult to

[†] See section 12.6.1 for contact addresses.

simply extract names, dates, places, etc. Most information retrieval systems still treat text as merely a bag of words with little or no linguistic structure. There have been numerous attempts to make use of richer linguistic structures such as phrases, predicate argument relations, and even morphology, but, thus far, most of these attempts have not resulted in significant improvements in retrieval performance.

Current natural language processing systems lack lexical and grammatical resources with sufficient coverage for unrestricted text. Consider the following famous pair of utterances:

> Time flies like an arrow.
> Fruit flies like a banana.

It would be useful for many applications to know that *fruit flies* is a phrase and *time flies* is not. Most systems currently do not have access to this kind of information. Parsers currently operate at the level of parts of speech, without looking at the words. Ultimately, parsers and other natural language applications will have to make greater use of collocational constraints and other constraints on words. The grammar/lexicon will have to be very large, at least as large as an 1800-page book (Quirk, Greenbaum, et al., 1985). The task may require a monumental effort like Murray's Oxford English Dictionary project.

Corpus-based methods may help speed up the lexical acquisition process by refining huge masses of corpus evidence into more manageable piles of high-grade ore. In Groliers encyclopedia (Grolier, 1991), for example, there are 21 instances of *fruit fly* and *fruit flies*, and not one instance of *time fly* and *time flies*. This kind of evidence is suggestive of the desired distinction, though far from conclusive.

12.2.3 Future Directions

The long-term research challenge is to derive lexicons and grammars for broad coverage natural language processing applications from corpus evidence.

A problem with attaining this long-term goal is that it is unclear whether the community of researchers can agree that a particular design of lexicons and grammars is appropriate, and that a large scale effort to implement that design will converge on results of fairly general utility (Liberman, 1992).

In the short-term, progress can be achieved by improving the infrastructure, i.e. the stock of intermediate resources mentioned in section 12.1. Data collection and dissemination efforts have been extremely successful. Efforts should now be focused on principles, procedures and tools for analyzing these data. There is a need for manual, semi-automatic and automatic methods that help produce linguistically motivated analyses that make it possible to derive further facts and generalizations that are useful in improving the performance of language processors.

While there is wide agreement in the research community on these general points, there seems to be no shared vision of what exactly to do with text corpora, once you have them. A way to procede in the short and intermediate term is for data collection efforts to achieve a consensus within the the research community by identifying a set of fruitful problems for research (e.g., word sense disambiguation, anaphoric reference, predicate argument structure) and collecting, analyzing and distributing relevant data in a timely and cost-effective manner. Funding agencies can contribute to the consensus building effort by encourging work on common tasks and sharing of common data and common components.

12.3 Spoken Language Corpora

Lori Lamel[a] & Ronald Cole[b]

[a] LIMSI-CNRS, Orsay, France

[b] Oregon Graduate Institute of Science & Technology, Portland, Oregon, USA

Spoken language is central to human communication and has significant links to both national identity and individual existence. The structure of spoken language is shaped by many factors. It is structured by the phonological, syntactic and prosodic structure of the language being spoken, by the acoustic enviroment and context in which it is produced—e.g., people speak differently in noisy or quiet environments—and the communication channel through which it travels.

Speech is produced differently by each speaker. Each utterance is produced by a unique vocal tract which assigns its own signature to the signal. Speakers of the same language have different dialects, accents and speaking rates. Their speech patterns are influenced by the physical environment, social context, the perceived social status of the participants, and their emotional and physical state.

Large amounts of annotated speech data are needed to model the affects of these different sources of variability on linguitic units such as phonemes, words, and sequences of words. An axiom of speech research is *there are no data like more data*. Annotated speech corpora are essential for progress in all areas of spoken language technology. Current recognition techniques require large amounts of training data to perform well on a given task. Speech synthesis systems require the study of large corpora to model natural intonation. Spoken languages systems require large corpora of human-machine conversations to model interactive dialogue.

In response to this need, there are major efforts underway worldwide to collect, annotate and distribute speech corpora in many languages. These corpora allow scientists to study, understand, and model the different sources of variability, and to develop, evaluate and compare speech technologies on a common basis.

Spoken Language Corpora Activities

Recent advances in speech and language recognition are due in part to the availability of large public domain speech corpora, which have enabled comparative system evaluation using shared testing protocols. The use of common corpora for developing and evaluating speech recognition algorithms is a fairly recent development. One of first corpora used for common evaluation, the TI-DIGITS corpus, recorded in 1984, has been (and still is) widely used as a test base for isolated and connected digit recognition (Leonard, 1984).

In the United States, the development of speech corpora has been funded mainly by agencies of the Department of Defense (DoD). Such DoD support produced two early corpora: Road Rally for studying word spotting, and the King Corpus, for studying speaker recognition. As part of its human language technology program, the Advanced Research Projects Agency (ARPA) of the DoD has funded TIMIT (Garofolo, Lamel, et al., 1993; Fisher, Doddington, et al., 1986; Lamel, Kassel, et al., 1986), a phonetically transcribed corpus of read sentences used for modeling phonetic variabilities and for evaluation of phonetic recognition algorithms, and task related corpora such as Resource Management (RM) (Price, Fisher, et al., 1988) and Wall Street Journal (WSJ) (Paul & Baker, 1992) for research on continuous speech recognition, and ATIS (Air Travel Information Service) (Price, 1990; Hirschmann, 1992) for research on spontaneous speech and natural language understanding.[1]

Recognition of the need for shared resources led to the creation of the Linguistic Data Consortium (LDC)[†] in the U.S. in 1992 to promote and support the widespread development and sharing of resources for human language technology. The LDC supports various corpus development activities, and distributes corpora obtained from a variety of sources. Currently, LDC distributes about twenty differerent speech corpora including those cited above, comprising many hundreds of hours of speech. Information about the LDC as well as contact information for most of the corpora mentioned below is listed in the next subsection.

The Center for Spoken Language Understanding (CSLU)[†] at the Oregon Graduate Institute collects, annotates and distributes telephone speech corpora. The Center's activities are supported by its industrial affiliates, but the corpora are made available to universities worldwide free of charge. Overviews of speech corpora available from the Center, and current corpus development activities, can be found in: Cole, Noel, et al. (1994); Cole, Fanty, et al. (1994). CSLU's Multi-Language Corpus (also available through the LDC), is the NIST standard for evaluating language identification algorithms, and is comprised of spontaneous speech in eleven different languages (Muthusamy, Cole, et al., 1992).

Europe is by nature multilingual, with each country having their own language(s), as well as dialectal variations and lesser used languages. Corpora development in Europe is thus the result of both National efforts and efforts sponsored by the European Union (typically under the ESPRIT (European Strategic Programme for Research and Development in Information Technology), LRE (Linguistic Research and Engineering), and TIDE (Technology Initiative for Disabled and Elderly People) programs, and now

---

[1] ARPA also sponsors evaluation tests, run by NIST (National Institute for Science and Technology), described in section 13.6.

[†] See section 12.6.2 for contact addresses.

for Eastern Europe under the PECO (Pays d'Europe Centrale et Orientalle)/Copernicus programs).

In February 1995 the European Language Resources Association (ELRA)[†] was established to provide a basis for central coordination of corpora creation, management and distribution in Europe. ELRA is the outcome of the combined efforts of partners in the LRE Relator[†] project and the LE MLAP (Language Engineering Multilingual Action Plan) projects: SPEECHDAT,[†] PAROLE[†] and POINTER.[†] These projects are responsible, respectively, for the infrastructure for spoken resources, written resources, and terminology within Europe. ELRA will work in close coordination with the Network of Excellence, ELSNET (European Network in Language and Speech),[†] whose Reusable Resources Task Group initiated the Relator project.

Several ESPRIT projects have attempted to create multilingual speech corpora in some or all of the official European languages. The first multilingual speech collection action in Europe was in 1989, consisting of comparable speech material recorded in five languages: Danish, Dutch, English, French, Italian. The entire corpus, now known as EUROM0 includes eight languages (Fourcin, Harland, et al., 1989). Other European projects producing corpora which may be available for distribution include: ACCOR[†] (multisensor recordings, seven languages, Marchal & Hardcastle, 1993); ARS;[†] EUROM1[†] (eleven languages); POLYGLOT[†] (seven languages LIMSI, 1994); ROARS;[†] SPELL;[†] SUNDIAL;[†] and SUNSTAR.[†]

The LRE ONOMASTICA[†] project (Trancoso, 1995) is producing large dictionaries of proper names and place names for eleven European languages. While some of these corpora are widely available, others have remained the property of the project consortium that created it. The LE SPEECHDAT project is recording comparable telephone data from 1000 speakers in eight European languages. A portion of the data will be validated and made publicly available for distribution by ELRA.

Some of the more important corpora in Europe resulting from National efforts are: **British English**: WSJCAM0[†] (Robinson, Fransen, et al., 1995), Bramshill,[†] SCRIBE,[†] and Normal Speech Corpus;[†] **Scotish English**: HCRC Map Task (Anderson, Bader, et al., 1991; Thompson, Anderson, et al., 1993); **Dutch**: Groningen;[†] **French**: BDSONS (Carré, Descout, et al., 1984), BREF[†] (Lamel, Gauvain, et al., 1991; Gauvain, Lamel, et al., 1990; Gauvain & Lamel, 1993); **German**: PHONDAT1 and PHONDAT2,[†] ERBA[†] and VERBMOBIL;[†] **Italian**: APASCI (Angelini, Brugnara, et al., 1993; Angelini, Brugnara, et al., 1994); **Spanish**: ALBAYZIN[†] (Moreno, Poch, et al., 1993; Diaz, Rubio, et al., 1993); **Swedish**: CAR and Waxholm.[†]

Some of these corpora are readily available (see the following section for contact information on corpora mentioned in this section); and efforts are underway to obtain the availability of others.

There have also been some recent efforts to record everyday speech of typical citizens. One such effort is part of the British National Corpus in which about 1500 hours of speech representing a demographic sampling of the population and wide range of materials has been recorded ensuring coverage of four contextual categories: educational, business, public/institutional, and leisure. The entire corpus is in the process of being orthographically transcribed with annotations for non-speech events. A similar corpus for Dutch is currently under discussion in the Netherlands, and the Institute of Phonetics and Verbal Communication of the University Munich has begun collecting of a very large database of spoken German.

The Translanguage English Database (TED) (Lamel, Schiel, et al., 1994) is a corpus of multi-dialect English and non-native English of recordings of oral presentations at Eurospeech'93 in Berlin. TEDspeeches contains data ranging in style from read to spontaneous, under varying degress of stress. An associated text corpus TEDtexts contains written versions of the proceedings articles, which can be used to define vocabulary items and to construct language models. Two auxilliary sets of recordings were made: one consisting of speakers recorded with a laryngograph (TEDlaryngo) in addition to the standard microphone, and the other a set of Polyphone-like recordings (TEDphone) made by the speakers in English and in their mother language. This corpus was partially funded by the LRE project EuroCocosda.[†]

Other major efforts in corpora collection have been undertaken in other parts of the world. These include: Polyphone, a multilingual, multinational application-oriented telephone speech corpus (co-sponsored by the LDC); the Australian National Database of Spoken Language (ANDOSL)[†] project, sponsored by the Australian Speech Science and Technology Association Inc. and funded by a research infrastructure grant from the Australian Research Council, is a national effort to create a database of spoken language; the Chinese National Speech Corpus[†] supported by the National Science Foundation of China designed to provide speech data for the acquisition of acoustic-phonetic knowledge and for the development and evaluation of speech processing systems; and corpora from Japan such as those publicly available from ATR, ETL and JEIDA.[†]

Future Directions

Challenges in spoken language corpora are many. One basic challenge is in design methodology—how to design compact corpora that can be used in a variety of applications; how to design comparable corpora in a variety of languages; how to select (or sample) speakers so as to have a representative population with regard to many factors including accent, dialect, and speaking style; how to create generic dialogue corpora so as to minimize the need for task or application specific data; how to select statistically representative test data for system evaluation. Another major challenge centers on developing standards for transcribing speech data at different levels and across languages: establishing symbol sets, alignment conventions, defining levels of transcription (acoustic, phonetic, phonemic, word and other levels), conventions for prosody and tone, conventions for quality control (such as having independent labelers transcribe the same speech data for reliability statistics). Quality control of the speech data is also an important issue that needs to be addressed, as well as methods for dissemination. While CDROM has become the defacto standard for dissemination of large corpora, other potential means need to also be considered, such as very high speed fiber optic networks.

12.4 Lexicons

Ralph Grishman[a] **& Nicoletta Calzolari**[b]

[a] New York University, New York, USA
[b] Istituto di Linguistica Computazionale del CNR, Pisa, Italy

12.4.1 The Lexicon as a Critical Resource

Lexical knowledge—knowledge about individual words in the language—is essential for all types of natural language processing. Developers of machine translation systems, which from the beginning have involved large vocabularies, have long recognized the lexicon as a critical (and perhaps the critical) system resource. As researchers and developers in other areas of natural language processing move from toy systems to systems which process real texts over broad subject domains, larger and richer lexicons will be needed and the task of lexicon design and development will become a more central aspect of any project. See Walker, Zampolli, et al. (1995); Zampolli, Calzolari, et al. (1994) for a rich overview of theoretical and practical issues connected with the lexicon in the last decade.

An important critical step towards avoiding duplication of efforts, and consequently towards a more productive course of action for the realization of resources, is to build and make publicly available to the community large-scale lexical resources, with broad coverage and basic types of information, generic enough to be reusable in different application frameworks, e.g., with application specific lexicons built on top of them. This need for shareable resources, possibly built in a cooperative way, brings in the issue of standardization and the necessity of agreeing on common/consensual specifications (Calzolari, 1994).

12.4.2 Types of Information

The lexicon may contain a wide range of word-specific information, depending on the structure and task of the natural language processing system. A basic lexicon will typically include information about morphology, either in a form enabling the generation of all potential word-forms associated with pertinent morphosyntactic features, or as a list of word-forms, or as a combination of the two. On the syntactic level, it will include in particular the complement structures of each word or word sense. A more complex lexicon may also include semantic information, such as a classification hierarchy and selectional patterns or case frames stated in terms of this hierarchy. For machine translation the lexicon will also have to record correspondences between lexical items in the source and target language; for speech understanding and generation it will have to include information about the pronunciation of individual words.

Strictly related to the types of information connected with each lexical entry are two other issues: (i) the overall lexicon architecture, and (ii) the representation formalism used to encode the data.

In general, a lexicon will be composed of different modules, corresponding to the different levels of linguistic descriptions, linked to each other according to the chosen overall architecture.

As for representation, we can mention at least two major formalisms. In an exchange model, Standard Generalized Markup Language (SGML) is widely accepted as a way of representing not only textual but also lexical data. The TEI (Text Encoding Initiative) has developed a model for representing machine readable dictionaries. In application systems, TFS (Typed Feature Structure) based formalisms are nowadays used in a large number of European lexical projects (Briscoe, Copestake, et al., 1993).

12.4.3 Sources of Information

Traditionally, computer lexicons have been built by hand specifically for the purpose of language analysis and generation. These lexicons, while they may have been large and expensive to build, have generally been crafted to the needs of individual systems and have not been treated as major resources to be shared among groups.

However, the needs for larger lexicons are now leading to efforts for the development of common lexical representations and co-operative lexicon development. They are leading developers to make greater use of existing resources—in particular, published commercial dictionaries—for automated language processing. And, most recently, the availability of large computer-readable text corpora has led to research on learning lexical characteristics from instances in text.

12.4.4 Major Projects

Among the first lexicons to be seen as shared resources for computational linguistics were the machine-readable versions of published dictionaries. One of the first major efforts involved a machine-readable version of selected information from Merriam-Webster's 7th Collegiate Dictionary, which was used for experiments in a number of systems. British dictionaries for English language learners have been especially rich in the information they encode—such as detailed information about complement structures—and so have proven particularly suitable for automated language processing. The Longman's Dictionary of Contemporary English, which included (in the machine-readable version) detailed syntactic and semantic codes, has been extensively used in computational linguistics systems (Boguraev & Briscoe, 1989); the Oxford Advanced Learner's Dictionary has also been widely used.

The major project having as its main objective the reuse of information extracted from Machine Readable Dictionaries (MRDs) is ESPRIT BRA (Basic Research Action) ACQUILEX. The feasibility of acquiring interesting syntactic/semantic information has been proved within ACQUILEX, using common extraction methodologies and techniques over more than ten MRDs in four languages. The objective was to build a prototype common Lexical Knowledge Base (LKB), using a unique Type System for all the languages and dictionaries, with a shared metalanguage of attributes and values.

Over the last few years there have been a number of projects to create large lexical resources for general use (see Varile & Zampolli, 1992 for an overview of international projects). The largest of these has been the Electronic Dictionary Research (EDR) project in Japan, which has created a suite of interlinked dictionaries, including Japanese and English dictionaries, a concept dictionary, and bilingual Japanese-English and English-Japanese dictionaries. The concept dictionary includes 400,000 concepts, both classified and separately described; the word dictionaries contain both grammatical information and links to the concept hierarchy.

In the United States, the WordNet Project at Princeton has created a large network of word senses related by semantic relations such as synonymy, part-whole, and is-a relations (Miller, 1990). The Linguistic Data Consortium (LDC) is sponsoring the creation of several lexical resources, including Comlex Syntax, an English lexicon with detailed syntactic information being developed at New York University.

Semantic Taxonomies similar or mappable to WordNet already exist (e.g., for Italian) or are being planned for a number of European languages, stemming from European projects.

The topic of large shareable resources has seen in the last years in Europe the flourishing of a number of important lexical projects, among which we can mention ET-7, ACQUILEX, ESPRIT MULTILEX, EUREKA GENELEX, MLAP ET-10 on Semantics acquisition from Cobuild, and LRE DELIS on corpus based lexicon development.

This concentration of efforts towards lexicon design and development in a multilingual setting has clearly shown that the area is ripe—at least for some levels of linguistic description—for reaching, in the short term, a consensus on common lexical specifications. The CEC DGXIII recently formed LRE EAGLES (Expert Advisory Group on Linguistic Engineering Standards) for pooling together the European efforts of both academic and industrial participants towards the creation of standards, among others in the lexical area (Calzolari & McNaught, 1994). A first proposal of common specifications at the morphosyntactic level has been prepared (Monachini & Calzolari, 1994), accompanied with language specific applications for the European languages.

12.4.5 Future Directions

Although there has been a great deal of discussion, design, and even development of lexical resources for shared use in computer analysis, there has been little practical experience with the actual use of such resources by multiple NLP projects. The sharing which has taken place has involved primarily basic syntactic information, such as parts of speech and basic subcategorization information; we have almost no experience with the sorts of semantic knowledge that could be effectively used by multiple systems. To gather such experience, we must provide ongoing support for several such lexical resources, and in particular provide support to modify them in response to users' needs.

We must also recognize the importance of the rapidly growing stock of machine-readable text as a resource for lexical research. There has been significant work on the discovery of subcategorization patterns and selectional patterns from text corpora. The major areas of potential results in the immediate future seem to lie in the combination of lexicon and corpus work. We see a growing interest from many groups in topics such as sense tagging or sense disambiguation on very large text corpora, where lexical tools and data provide a first input to the systems and are in turn enhanced with the information acquired and extracted from corpus analysis.

12.5 Terminology[2]

Christian Galinski[a] & Gerhard Budin[b]

[a] Infoterm, Vienna, Austria
[b] Univerity of Vienna, Austria

12.5.1 What is Terminology?

Whenever and wherever specialized information and knowledge are created, communicated, recorded, processed, stored, transformed or re-used, terminology is involved in one way or another. Subject-field communication has become a specific type of discourse with specialized texts differentiating into a whole array of text types. When we define terminology as a structured set of concepts and their designations in a particular subject field, it can be considered the infrastructure of specialized knowledge. Technical writing and technical documentation are thus impossible without properly using terminological resources. Since the production of technical texts increasingly involves several languages, high-quality multilingual terminologies have become scarce and much desired commodities on the burgeoning markets of language and knowledge industries.

12.5.2 Interdisciplinary Research

The research field we talk about is referred to as terminology science, its practical field of application is in terminology management, which includes the creation of subject-field specific terminologies and the terminographic recording of such information in the form of terminology databases, dictionaries, lexicons, specialized encyclopedias, etc. (For overviews and recent textbooks see, for English: Felber, 1984; Picht & Draskau, 1985; Sager, 1990; for German: Felber & Budin, 1989; Arntz & Picht, 1989; for Spanish: Cabré, 1994; for French: Gouadec, 1992.)

Concepts are considered the smallest units (*atoms*) of specialized knowledge. They never occur in isolation, but rather in complex conceptual networks that are multidimensional, due to a wide range of conceptual relationships among concepts. Given the limitations of natural language with regard to the representation of these concepts in specialized discourse (limited number of term elements in every language), concepts are increasingly

[2] This section has been compiled on the basis of current discussions in terminology science and experience in a multitude of terminological activities world-wide.

represented by non-linguistic designations, like graphical symbols (Galinski & Picht, 1995). In addition we may distinguish between:

- symbolic representations
  - terms (including abbreviations, alphanumeric codes, etc.)
  - graphical symbols, audiovisual symbols, etc.
  - combinations of both
- descriptive representations
  - definitions, explanations, etc. as linguistic descriptions of concepts
  - pictures, charts, graphics, etc. as graphical/pictorial descriptions of concepts
  - combinations of both

Theories of terminology as they have developed over at least six decades, consider concepts as:

- **units of thought**, focusing on the psychological aspect of recognizing objects as part of reality;
- **units of knowledge**, focusing on the epistemological aspect of information gathered (today we say constructed) on the object in question;
- **units of communication**, stressing the fact that concepts are the prerequisite for knowledge transfer in specialized discourse (Galinski, 1990).

The development of terminologies as a crucial part of special purpose languages reflects scientific, technical and economic progress in the subject fields concerned. Due to different speeds in this dynamic co-evolution of knowledge in the individual domains, specialized discourse continues to differentiate into more and more sectorized special languages and terminologies. But these communication tools become increasingly ambiguous, due to the sheer number of concepts to be designated and the limited linguistic resources of every natural language: terms are taken over from one domain (or language) into another, usually with varying meanings in the (productive) form of metaphors or analogies; new homonyms, polysemes and synonyms arise, motivating or even forcing subject specialists to standardize their terminology and harmonize them on the multilingual level in order to reduce and manage the constantly rising communicative complexity that faces their discourse communities.

But terminology research is not limited to comparative semiotic and linguistic studies of term formation and the epistemological dimension of the evolution of scientific knowledge. The agenda of terminology science also includes socio-terminological studies of the acceptance of neologisms proposed by terminology and language planners (Gaudin, 1994), case studies on terminology development by standardization and harmonization efforts, research and development concerning the establishment and use of terminology databases for various user groups and purposes (e.g., translation, technical writing, information management) and concerning controlled vocabularies for documentation and information retrieval purposes (thesauri, classification systems, etc.).

12.5.3 Terminology Management

Terminology management is primarily concerned with manipulating terminological resources for specific purposes, e.g., establishing repertories of terminological resources for publishing dictionaries, maintaining terminology databases, or ad hoc problem solving in finding multilingual equivalences in translation work or creating new terms in technical writing. (For terminology management see Wright & Budin, 1995.)

Terminology databases are increasingly available by on-line query or on CD-ROM (e.g., TERMIUM, EURODICAUTOM), on diskette in the form of electronic dictionaries or as private databases established and maintained by engineers, computer specialists, chemists, etc. (working as terminologists, translators, technical writers) for various purposes:

(a) computer-assisted human translation;

(b) computer-assisted technical and scientific writing;

(c) materials information systems (spare parts administration, etc.);

(d) terminology research in linguistics, information science, philosophy of science, sociology of technology, etc.

For such purposes special computer programs have been developed (terminology database management programs), either commercially available on the international terminology market or developed as prototypes in academic research projects.

Due to the surprisingly high diversity of terminological resources that is potentially relevant to applications, terminology databases may look quite different from each other. One principle, however, seems to be the common denominator of all of them: the concepts under consideration are always the point of departure for database modeling; entries in terminology databases deal with one specific concept at a time. A terminological entry may contain not only term equivalents in other languages, synonyms, abbreviations, regional variants, definitions, contexts, even graphics or pictures, but also indications of relationships to other concepts (referencing to related entries) and subject-field indications by including thesaurus descriptors, class names from a classification system, etc., in order to easily retrieve terminological entries covering a certain topic.

12.5.4 Future Directions

Theoretical Issues: The last few years have seen a considerable increase in epistemological studies in the framework of philosophy of science concerning the way in which scientific knowledge is constantly created, communicated and changed and the pivotal role scientific terminologies play in this respect. In the light of post-modernism, complexity, fractal and chaos theories, synergetics and other new paradigms that completely change our scientific view of the world and of ourselves, it is necessary to re-examine the correspondence between *objects* we perceive or conceive and the concepts we construct in the process of thinking (cognition, and re-cognition of objects) (De Beaugrande, 1988; Budin, 1994).

The concept-oriented approach in terminology management mentioned above seems to be the key to solve a whole range of methodological problems in the management of multilingualism and information management in large international institutions as a number of innovate projects on the European level could prove (Galinski, 1994). The performance of machine translation systems system could also be improved by integrating advanced terminology processing modules that are based on the conceptual approach to language engineering.

European research projects such as Translator's Workbench (ESPRIT Programme) or similar projects in Canada (e.g., Translation Workstation), show a clear tendency towards systems integration: terminology products are no longer isolated and difficult to use, but fully integrated in complex work environments. Automatic term extraction from text corpora is one of the buzzwords in this type of practice-oriented research (Ahmad, Davies, et al., 1994). A terminological analysis of text corpora also includes fuzzy matching in order to recognize larger segments of texts (complex multi-word terms, fixed collocations, but also semi-fixed sentence patterns).

Within the framework of the Text Encoding Initiative a working group (i.e., TEI A&I-7) has specifically been devoted to terminological resources and their management by SGML. Chapter 13 of the P 2 Guidelines of TEI on the application of SGML in text processing is dealing with the representation of terminological resources in SGML and the creation of an interchange format. This terminology interchange format (TIF) is now in the process of being standardized by ISO (ISO 12200, Melby, Budin, et al., 1993). The exchange of terminological resources has become one of the most discussed topics in the international terminology community. In addition to the introduction of the TIF standard, many methodological and legal problems (copyright, intellectual property rights, etc.) have to be solved.

Terminologists have also joined the international bandwagon of quality assurance and total quality management by starting research projects on how appropriate terminology management may improve the performance of quality managers, and *vice versa*, how to improve reliability of terminological resources by systematic quality management in terminology standardization in particular and terminology management in general.

The interdisciplinary nature of terminology science also becomes clear in its links to research in knowledge engineering and Artificial Intelligence Research (Ahmad, 1995; Schmitz, 1993). But terminological knowledge engineering (TKE) is more than just a series of projects and some new tools—it has also become a new method of modeling and representing knowledge in hypermedia knowledge bases serving as research tools for scientists (and completely changing research methods, e.g., by terminology visualization modules), as a knowledge popularization tool in museums, and as a teaching tool or as a *hyperterminology* database (IITF, 1994).

12.6 Addresses for Language Resources

12.6.1 Written Language Corpora

Contact information for the corpora mentioned in section 12.2 is provided here in alphabetical order.

British National Corpus (BNC): smbowie@vax.oxford.ac.uk

Consortium for Lexical Research (CLR): lexical@nmsu.edu

Dansk Korpus (DK): olenc@coco.ihi.ku.dk (Ole Norling-Christensen)

European Corpus Initiative (ECI): (in Europe): eucorp@cogsci.edinburgh.ac.uk

European Corpus Initiative (ECI): (in U.S.) LDC: ehodas@unagi.cis.upenn.edu

Frantext of Institut National de la Langue Francaise (INaLF-CNRS): emartin@FRCII171 (Eveline Martin)

Institut für deutsche Sprache (IDS): neumann@ids-mannheim.de (Robert Neumann)

Instituut voor Nederlandse Lexicologie (INL): postmaster@hnympi52.bitnet

International Computer Archive of Modern English (ICAME): stigj@hedda.uio.no (Stig Johansson)

Istituto di Linguistica Computazionale (ILC-CNR): glottolo@vm.cnuce.cnr.it (Antonio Zampolli)

Linguistic Data Consortium (LDC): ehodas@unagi.cis.upenn.edu (Elizabeth Hodas). The WWW page is *http://www.cis.upenn.edu/ ldc*. Information about the LDC and its activities can also be obtained via anonymous FTP *ftp.cis.upenn.edu* under *pub/ldc*. Most of the data are compressed using the tool Shorten by T. Robinson which is available via ft *svr-ftp.eng.cam.ac.uk*

Norsk Tekstarkiv: per.vestbostad@hd.uib.no (Per Vestbostad)

Spanish Reference Corpus Project: marcos@emduam11.bitnet (Francisco Marcos Marin), Sociedad Estatal del V Centenario

Stockholm-Umea Corpus (SUC): gunnel@ling.su.se (Gunnel Kallgren); ejerhed@ling.umu.se (Eva Ejerhed); Sprakdata gellerstam@svenska.gu.se (Martin Gellerstam)

Text Encoding Initiative (TEI): lou@vax.ox.ac.uk (Lou Burnard), u35395@uicvm.bitnet (C.M. Sperberg McQueen)

University of Helsinki: fkarlsso@ling.helsinki.fi (Fred Karlsson)

12.6.2 Spoken Language Corpora

Contact information for the corpora mentioned in section 12.3 is provided here in alphabetical order.

ACCOR: Project contact: Prof. W. Hardcastle, *sphard@queen-margaret-college.main.ac.uk*; Prof. A. Marchal, *phonetic@fraix11.bitnet* (The British English portion of the ACCOR corpus is being produced on CDROM with partial financing from ELSNET)

ALBAYZIN: Corpus contact: Professor Climent Nadeu, Department of Speech Signal Theory and Communications, Universitat Politecnica de Catalunya, ETSET, Apartat 30002, 08071 Barcelona, Spain, *nadeu@tsc.upc.es*

ARS: CSELT (coordinator), Mr. G. Babini, Via G. Beis Romoli 274, I-101488, Torino, Italy

ATR, ETL & JEIDA: Contact person: K. Kataoka, AI and Fuzzy Promotion Center, Japan Information Processing Development Center (JIPDEC), 3-5-8 Shibakoen, Minatoku, Tokyo 105, Japan, TEL. +81 3 3432 9390, FAX. +81 3 3431 4324

Australian National Database of Spoken Language (ANDOSL): Corpus contact: Bruce Millar, Computer Sciences Laboratory, Research School of Information Sciences and Engineering, Australian National University, Canberra, ACT 0200, Australia, email: *bruce@cslab.anu.edu.au*

BREF: Corpus contact: send email to *bref@limsi.fr*

Bramshill: LDC (as above)

CAR & Waxholm: Corpus contact: Bjorn Granstrom *bjorn@speech.kth.se*

Center for Spoken Language Understanding (CSLU): Information on the collection and availability of CSLU corpora can be obtained on the World Wide Web, *http://www.cse.ogi.edu/CSLU/corpora.html*

Chinese National Speech Corpus: Contact person: Prof. Jialu Zhang, Academia Sinica, Institute of Acoustics, 17 Shongguanjun St, Beijing PO Box 2712, 100080 Beijing, Peoples Republic of China

ERBA: Corpus contact: Stefan Rieck, Lehrstuhl Informatik 5 (Pattern Recognition), University of Erlangen-Nurnberg, Martensstr.3 , 8520 Erlangen, Germany, Email: *rieck@informatik.uni-erlangen.de*

ETL: see ATR above.

EUROM1: Project contact for Multilingual speech database: A. Fourcin (UCL) *adrian@phonetics.ucl.ac.uk*; or the following for individual languages:
D: D. Gibbon (Un.Bielefeld) *gibbon@asl.uni-bielefeld.de*
DK: B. Lindberg (IES) *bli@stc.auc.dk*
F: J.F. Serignat (ICP) *serignat@icp.grenet.fr*
I: G. Castagneri (CSELT) *castagneri@cselt.stet.it*
N: T. Svendsen (SINTEF-DELAB) *torbjorn@telesun.tele.unit.no*
NL: J. Hendriks or L. Boves (PTT Research) *boves@lett.kun.nl*
SW: G. Hult (Televerket) or B. Granstrom (KTH) *bjorn@speech.kth.se*
UK: A. Fourcin (UCL) *adrian@phonetics.ucl.ac.uk*
Contact for SAM-A EUROM1:
E: A. Moreno (UPC) *amoreno@tsc.upc.es*
G: J. Mourjopoulos (UPatras) *mourjop@grpafvx1.earn*
P: I. Trancoso (INESC) *imt@inesc.pt*

EuroCocosda: Corpus contact: A Fourcin, email: *adrian@phonetics.ucl.ac.uk*

European Language Resources Association (ELRA): For membership information contact: Sarah Houston, email: *100126.1262@compuserve.com*

European Network in Language and Speech (ELSNET): OTS, Utrecht University, Trans 10, 3512 JK, Utrecht, The Netherlands, Email: *elsnet@let.ruu.nl*

Groningen: Corpus contact: Els den Os, Speech Processing Expertise Centre, P.O.Box 421, 2260 AK Leidschendam, The Netherlands, *els@spex.nl* (CDs available via ELSNET)

JEIDA: see ATR above.

LRE ONOMASTICA: Project contact: M. Jack, CCIR, University of Edinburgh, *mervyn.jack@ed.ac.uk*

Linguistic Data Consortium (LDC): see LDC above.

Normal Speech Corpus: Corpus Contact: Steve Crowdy, Longman UK, Burnt Mill, Harlow, CM20 2JE, UK

Oregon Graduate Institute (OGI): see CSLU above.

PAROLE: Project contact: Mr. T. Schneider, Sietec Systemtechnik Gmbh, Nonnendammallee 101, D-13629 Berlin

PHONDAT2: Corpus contact: B. Eisen, University of Munich, Germany

POINTER: Project contact: Mr. Corentin Roulin , BJL Consult, Boulevard du Souverain 207/12, B-1160 Bruxelles

POLYGLOT: Contact person: Antonio Cantatore, Syntax Sistemi Software, Via G. Fanelli 206/16, I- 70125 Bari, Italy

Relator: Project contact: A. Zampolli, Istituto di Linguistica Computazionale, CNR, Pisa, I, E-mail: *giulia@icnucevm.cnuce.cnr.it*; Information as well as a list of resources, is available on the World Wide Web, *http://www.XX.relator.research.ec.org*

ROARS: Contact person: Pierre Alinat, Thomson-CSF/Sintra-ASM, 525 Route des Dolines, Parc de Sophia Antipolis, BP 138, F-06561 Valbonne, France

SCRIBE: Corpus contact: Mike Tomlinson, Speech Research Unit, DRA, Malvern, Worc WR14 3PS, England

SPEECHDAT: Project contact: Mr. Harald Hoege, Siemens AG, Otto Hahn Ring 6, D-81739 Munich

SPELL: Contact person: Jean-Paul Lefevre, Agora Conseil, 185, Hameau de Chateau, F-38360 Sassenage, France

SUNDIAL: Contact person: Jeremy Peckham, Vocalis Ltd., Chaston House, Mill Court, Great Shelford, Cambs CB2 5LD UK, email: *jeremy@vocalis.demon.co.uk*

SUNSTAR: Joachin Irion, EG Electrocom Gmbh, Max-Stromeyerstr. 160, D- 7750 Konstanz, Germany

VERBMOBIL: Corpus contact: B. Eisen, University of Munich, Germany

Wall Street Journal, Cambridge, zero (WSJCAM0): Corpus contact: Linguistic Data Consortium (LDC), Univ. of Pennsylvania, 441 Williams Hall, Philadelphia, PA, USA 19104-6305, (215) 898-0464

Waxholm: see CAR above.

12.6.3 Character Recognition

Contact information for the corpora mentioned in section 13.10 is provided here in alphabetical order.

Electrotechnical Laboratory (ETL) Character Database: Distributor: Image Understanding Section, Electrotechnical Laboratory, 1-1-4, Umezono, Tsukuba, Ibaraki, 305, Japan.

National Institute of Standards and Technology (NIST): Distributor: Standard Reference Data, National Institute of Standards and Technology, 221/A323, Gaithersburg, MD 20899, USA.

U.S. Postal Service: Distributor: CEDAR, SUNY at Buffalo, Dept. of Computer Science, 226 Bell Hall, Buffalo, NY 14260, USA.

University of Washington: Distributor: Intelligent Systems Laboratory, Dept. of Electrical Engineering, FT-10, University of Washington, Seattle, WA 98195, USA

12.7 Chapter References

Ahmad, K. (1995). The analysis of text corpora for the creation of advanced terminology databases. In Wright, S. E. and Budin, G., editors, *The Handbook of Terminology Management*. John Benjamins, Amsterdam/Philadelphia.

Ahmad, K., Davies, A., Fulford, H., Holmes-Higgin, P., and Rogers, M. (1994). Creating terminology resources. In Kugler, M., Ahmad, K., and Thurmair, G., editors, *Research Reports ESPRIT: Translator's Workbench—Tools and Terminology and Text Processing, Project 2315 TWB*, volume 1, pages 59–71. Springer, Heidelberg, Berlin, New York.

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E. H., Doherty, G. M., Garrod, S. C., Isard, S. D., Kowtko, J. C., McAllister, J. M., Miller, J., Sotillo, C. F., Thompson, H. S., and Weinert, R. (1991). The HCRC map task corpus. *Language and Speech*, 34(4).

Angelini, B., Brugnara, F., Falavigna, D., Giuliani, D., Gretter, R., and Omologo, M. (1993). A baseline of a speaker independent continuous speech recognizer of Italian. In *Eurospeech '93, Proceedings of the Third European Conference on Speech Communication and Technology*, volume 2, pages 847–850, Berlin. European Speech Communication Association.

Angelini, B., Brugnara, F., Falavigna, D., Giuliani, D., Gretter, R., and Omologo, M. (1994). Speaker independent continuous speech recognition using an acoustic-phonetic Italian corpus. In *Proceedings of the 1994 International Conference on Spoken Language Processing*, volume 3, pages 1391–1394, Yokohama, Japan.

Arntz, R. and Picht, H. (1989). *Einführung in die übersetzungsbezogene terminologiearbeit*. Hildesheim, Zürich, New York.

ARPA (1993). *Proceedings of the 1993 ARPA Human Language Technology Workshop*, Princeton, New Jersey. Advanced Research Projects Agency, Morgan Kaufmann.

ARPA (1994). *Proceedings of the 1994 ARPA Human Language Technology Workshop*, Princeton, New Jersey. Advanced Research Projects Agency, Morgan Kaufmann.

Boguraev, B. and Briscoe, T., editors (1989). *Computational Lexicography for Natural Language Processing*. Longman.

Briscoe, T., Copestake, A., and de Pavia, V., editors (1993). *Inheritance, defaults and the lexicon*. Cambridge University Press.

Budin, G. (1994). Organisation und evolution von fachwissen und fachsprachen am beispiel der rechtswissenschaft [organization and evolution of specialized knowledge and specialized languages in the case of the law]. In Wilske, D., editor, *Erikoiskielet ja Käännösteoria [LSP and Theory of Translation]. VAKKI-symposiumi XIV [14th VAKKI Symposium]*, pages 9–21.

Cabré, T. (1994). *La terminologia*. Barcelona.

Calzolari, N. (1994). European efforts towards standardizing language resources. In Steffens, P., editor, *Machine Translation and the Lexicon*. Springer-Verlag.

Calzolari, N. and McNaught, J. (1994). EAGLES editors' introduction. Technical report, EAGLES Draft Editorial Board Report, EAGLES Secretariat, Istituto di Linguistica Computazionale, Via della Faggiola 32, Pisa, Italy 56126, Fax: +39 50 589055, E-mail: ceditor@tnos.ilc.pi.cnr.it.

Carré, R., Descout, R., Eskénazi, M., Mariani, J., and Rossi, M. (1984). The French language database: defining, planning, and recording a large database. In *Proceedings of the 1984 International Conference on Acoustics, Speech, and Signal Processing*. Institute of Electrical and Electronic Engineers.

Cole, R. A., Fanty, M., Noel, M., and Lander, T. (1994). Telephone speech corpus development at cslu. In *Proceedings of the 1994 International Conference on Spoken Language Processing*, volume 4, pages 1815–1818, Yokohama, Japan.

Cole, R. A., Noel, M., Burnett, D. C., Fanty, M., Lander, T., Oshika, B., and Sutton, S. (1994). Corpus development activities at the center for spoken language understanding. In *Proceedings of the 1994 ARPA Human Language Technology Workshop*, Princeton, New Jersey. Advanced Research Projects Agency, Morgan Kaufmann.

DARPA (1986). *Proceedings of the DARPA Speech Recognition Workshop*. Defense Advanced Research Projects Agency. SAIC-86/1546.

DARPA (1991). *Proceedings of the Third Message Understanding Conference*, San Diego, California. Morgan Kaufmann.

DARPA (1992). *Proceedings of the Fifth DARPA Speech and Natural Language Workshop*. Defense Advanced Research Projects Agency, Morgan Kaufmann.

De Beaugrande, R. (1988). Systemic versus contextual aspects of terminology. In *Terminology and Knowledge Engineering, Supplement*, pages 7–24. Indeks, Frankfurt.

Diaz, J., Rubio, A., Peinado, A., Segarra, E., Prieto, N., and Casacuberta, F. (1993). Development of task-oriented Spanish speech corpora. The paper was distributed at the conference and does not appear in the proceedings.

Eurospeech (1993). *Eurospeech '93, Proceedings of the Third European Conference on Speech Communication and Technology*, Berlin. European Speech Communication Association.

Felber, H. (1984). *Terminology Manual.* UNESCO, Paris.

Felber, H. and Budin, G. (1989). *Terminology Manual.* UNESCO, Paris, second edition.

Fisher, W., Doddington, G., and Goudie-Marshall, K. (1986). The DARPA speech recognition research database: Specifications and status. In *Proceedings of the DARPA Speech Recognition Workshop*, pages 93–99. Defense Advanced Research Projects Agency. SAIC-86/1546.

Fourcin, A. J., Harland, G., Barry, W., and Hazan, V. (1989). *Speech input and output assessment; multilingual methods and standards.* Ellis Horwood.

Francis, W. and Kucera, H. (1982). *Frequency Analysis of English Usage.* Houghton Mifflin, Boston.

Galinski, C. (1990). Terminology 1990. *TermNet News*, 1994(24):14–15.

Galinski, C. (1994). Terminologisches informationsmanagement in harmonisierungsprojekten der EU. Unpublished.

Galinski, C. and Picht, H. (1995). Graphic and other semiotic forms of knowledge representation in terminology work. In Wright, S. E. and Budin, G., editors, *Handbook of Terminology Management.* John Benjamins, Amsterdam/Philadelphia. winter 1995.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., and Dahlgren, N. L. (1993). The DARPA TIMIT acoustic-phonetic continuous speech corpus. CDROM: NTIS order number PB91-100354.

Gaudin (1994). Socioterminologie.

Gauvain, J.-L. and Lamel, L. F. (1993). Sous-corpus BREF 80, disques bref 80-1 et bref 80-2 (CDROM).

Gauvain, J.-L., Lamel, L. F., and Eskénazi, M. (1990). Design considerations & text selection for BREF, a large French read-speech corpus. In *Proceedings of the 1990 International Conference on Spoken Language Processing*, volume 2, pages 1097–1100, Kobe, Japan.

Gouadec, D. (1992). *La terminologie.* Afnor, Paris.

Grolier (1991). *New Grolier's Electronic Encyclopedia.* Grolier.

Hirschmann, L. (1992). Multi-site data collection for a spoken language corpus. In *Proceedings of the Fifth DARPA Speech and Natural Language Workshop.* Defense Advanced Research Projects Agency, Morgan Kaufmann.

ICASSP (1984). *Proceedings of the 1984 International Conference on Acoustics, Speech, and Signal Processing.* Institute of Electrical and Electronic Engineers.

ICSLP (1994). *Proceedings of the 1994 International Conference on Spoken Language Processing*, Yokohama, Japan.

IITF (1994). Final report. Multimedia knowledge database for social anthropology.

Lamel, L. F., Gauvain, J.-L., and Eskénazi, M. (1991). BREF, a large vocabulary spoken corpus for French. In *Eurospeech '91, Proceedings of the Second European Conference on Speech Communication and Technology*, Genova, Italy. European Speech Communication Association.

Lamel, L. F., Kassel, R. H., and Seneff, S. (1986). Speech database development: Design and analysis of the acoustic-phonetic corpus. In *Proceedings of the DARPA Speech Recognition Workshop*, pages 100–109. Defense Advanced Research Projects Agency. SAIC-86/1546.

Lamel, L. F., Schiel, F., Fourcin, A., Mariani, J., and Tillmann, H. G. (1994). The translanguage English database (TED). In *Proceedings of the 1994 International Conference on Spoken Language Processing*, volume 4, pages 1795–1798, Yokohama, Japan.

Leonard, R. G. (1984). A database for speaker-independent digit recognition. In *Proceedings of the 1984 International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 42.11–14. Institute of Electrical and Electronic Engineers.

Liberman, M. (1992). Core NL lexicons and grammars. In *Proceedings of the Fifth DARPA Speech and Natural Language Workshop*, page 351 (session 10b). Defense Advanced Research Projects Agency, Morgan Kaufmann.

LIMSI (1994). Sous-corpus BREF polyglot (CDROM).

Marchal, A. and Hardcastle, W. J. (1993). ACCOR: Instrumentation and database for the cross-language study of coarticulation. *Language and Speech*, 36:137–153.

Melby, A., Budin, G., and Wright, S. E. (1993). The terminology interchange format (TIF)—a tutorial. *TermNet News*, 1993(40):9–65.

Miller, G. (1990). Wordnet: An on-line lexical database. *International journal of Lexicography*, 3(4):235–312.

Monachini, M. and Calzolari, N. (1994). Synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora and applications to european languages. Technical report, EAGLES ILC Pisa, EAGLES Secretariat, Istituto di Linguistica Computazionale, Via della Faggiola 32, Pisa, Italy 56126, Fax: +39 50 589055, E-mail: ceditor@tnos.ilc.pi.cnr.it. Also available via ftp to nicolet.ilc.pi.cnr.it (131.114.41.11), Username: eagles, Password: eagles.

Moreno, A., Poch, D., Bonafonte, A., Lleida, E., Llisterri, J. R., Marino, J. B., and Nadeu, C. (1993). ALBAYZIN speech database: design of the phonetic corpus. In *Eurospeech '93, Proceedings of the Third European Conference on Speech Communication and Technology*, volume 1, pages 175–178, Berlin. European Speech Communication Association.

Muthusamy, Y. K., Cole, R. A., and Oshika, B. T. (1992). The OGI multi-language telephone speech corpus. In *Proceedings of the 1992 International Conference on Spoken Language Processing*, volume 2, pages 895–898, Banff, Alberta, Canada. University of Alberta.

Paul, D. and Baker, J. (1992). The design for the Wall Street Journal-based CSR corpus. In *Proceedings of the Fifth DARPA Speech and Natural Language Workshop*. Defense Advanced Research Projects Agency, Morgan Kaufmann.

Picht, H. and Draskau, J. (1985). *Terminology, An Introduction*. The Copenhagen School of Economics, Copenhagen.

Price, P. (1990). Evaluation of spoken language systems: The ATIS domain. In *Proceedings of the Third DARPA Speech and Natural Language Workshop*, Hidden Valley, Pennsylvania. Defense Advanced Research Projects Agency, Morgan Kaufmann.

Price, P., Fisher, W. M., Bernstein, J., and Pallett, D. S. (1988). The DARPA 1000-word resource management database for continuous speech recognition. In *Proceedings of the 1988 International Conference on Acoustics, Speech, and Signal Processing*, pages 651–654, New York. Institute of Electrical and Electronic Engineers.

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. Longman.

Robinson, T., Fransen, J., Pye, D., Foote, J., and Renals, S. (1995). WSJCAM0: A british english speech corpus for large vocabulary continuous speech recognition. In *Proceedings of the 1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 81–84, Detroit. Institute of Electrical and Electronic Engineers.

Sager, J. C. (1990). *A practical course in terminology processing.* John Benjamins, Amsterdam/Philadelphia.

Schmitz, K.-D. (1993). TKE 93. terminology and knowledge engineering. In Schmitz, K.-D., editor, *Proceedings of the 3rd International Congress*, Cologne, Germany. Frankfurt a.M.: INDEKS Verlag.

Thompson, H. S., Anderson, A., Bard, E. G., Boyle, E. H., Doherty-Sneddon, G., Newlands, A., and Sotillo, C. (1993). the HCRC map task corpus: Natural dialog for speech recognition. In *Proceedings of the 1993 ARPA Human Language Technology Workshop*, Princeton, New Jersey. Advanced Research Projects Agency, Morgan Kaufmann.

Trancoso, I. (1995). The onomastica inter-language pronunciation lexicon. In *Eurospeech '95, Proceedings of the Fourth European Conference on Speech Communication and Technology*, Madrid, Spain. European Speech Communication Association. In press.

Varile, N. and Zampolli, A., editors (1992). *COLING92 International Project Day.* Giardini Editori, Pisa.

Walker, D., Zampolli, A., and Calzolari, N., editors (1995). *Automating the Lexicon: Research and Practice in a Multilingual Environment.* Oxford University Press.

Wright, S. E. and Budin, G., editors (1995). *Handbook of Terminology Management.* John Benjamins, Amsterdam/Philadelphia. winter 1995.

Zampolli, A., Calzolari, N., and Palmer, M., editors (1994). *Current Issues in Computational Linguistics: In Honour of Don Walker.* Giardini Editori, Pisa and Kluwer, Dordrecht.