The ELRA Newsletter

February 1998

Vol.3 n.1

Table of contents

Signed articles represent the views of their authors and do not necessarily reflect the position of the Editors, or the official policy of the ELRA Board/ELDA staff.

ISSN: 1026-8200

ELRA/ELDA
CEO: Khalid Choukri
Assistant: Rébecca Jaffrain

55-57 rue Brillat Savarin
75013 Paris
Phone: (33) 1 43 13 33 33
Fax: (33) 1 43 13 33 30
E-mail:
elra@calvanet.calvacom.fr
WWW:
http://www.icp.grenet.fr/ELRA/home.html

Dear ELRA Members,

The beginning of 1998 has found ELRA hard at work at its core business of acquiring and distributing language resources, we have since December distributed some 16 speech resources and 13 text resources. The new resources featured in this issue are the Bilingual Collocational Dictionary from Horst Bogatz and the SPK speech database from ITC-IRST in Italy.

In addition, the corpus and lexicon validation manuals are now both available from the ELDA office and on the ELRA Web site. The market study has been completed and the full results will be sent to members and other survey participants shortly. We are also proud to welcome four new members in 1998: Daimler-Benz Aerospace, Germany; Wildfire Communications Inc., USA; Sony International, Germany and Texas Instruments Inc., USA.

ELRA has increased its project activities. Work has included attending the kick-off meeting of ELSE, which will address the problems of an evaluation infrastructure for speech and language in Europe. One question to be answered in this context is whether ELRA should become a centre for evaluating technologies/applications or whether it should simply provide resources for evaluation to a third party (e.g. a new infrastructure or association or a network of evaluation and assessment sites). We have also prepared a number of documents for the SALA (SpeechDat Across Latin America) project, including draft co-production agreements. ELRA has participated in an ESPRIT proposal on the reuse of ELRA resources in the translation of keywords in Web pages, and has participated in a road-map discussion with ELSNET. Last but not least, ELRA will join the interest group for the SENSEVAL project, which is working toward the semantic tagging of a corpus to be used for tagger evaluation. ELRA proposes to supply the raw data and distribute the tagged data.

In the mean time, we have still found time to relocate to new offices (see the address on the cover of this issue for full details) with effect from February 1. This move enables us to expand our activities according to the plans for 1998 and on top of that welcome our board and committees for meetings in more convenient premises.

We have also been pushing ahead with preparations for the LREC conference in Granada at the end of May. Over 200 papers have been accepted, along with eight workshop proposals. We are looking forward to what will obviously be a major event in language engineering in Europe. In parallel with the conference, ELRA will be organising an exhibition for companies or projects wishing to demonstrate their products or prototypes. For more details, contact the ELDA office or send an e-mail to elra-elda@calva.net. If you have not already registered for the Conference, we urge you to do so now. You will find the registration form on the ELRA Web site (http://www.icp.grenet.fr/ELRA/home.html). Remember that all ELRA members are granted registration of one person free of charge!

In this issue, there are a number of valuable articles on different aspects of language engineering. Melvyn Hunt of Dragon Systems UK has provided a clear overview of the state of the art in dictation systems, while Frédérique Segond and colleagues at Rank Xerox have reported on their development of multilingual NLP systems. Other reports feature EuroWordnet (Piek Vossen) and the APOLLO project (Guy Deville and Pierre Mousel). In addition, ELRA is proud to present evidence of the growing recognition of this organisation outside the immediate language community, in the form of an article by Bernard Montelh which appeared in *Le Monde* on 1 February 1998, in which the ELRA CEO expressed his view on the importance of language resources and in the French Government's action programme on "Preparing France's Entry into the Inform@tion Society". After the hard work of our set-up period, it is gratifying to note how the visibility and appreciation of our organisation, and hence of you, its members, is now obviously rising.

Antonio Zampolli, President Khalid Choukri, CEO

Erratum:

In the article by Florian Schiel, "Probabilistic analysis of pronounciation with MAUS" (Newsletter Vol.2 N.4, p.6-9), a section of the text describing evaluation of a phonetic/phonemic segmentation of arbitrary utterances, was left out. We wish to express our apologies to the author and our readers for this unfortunate oversight. The paragraph missing is enclosed as a separate page.

Latest News on the LREC

28-30 May 1998 in Granada, Spain

The preparations for the Granada conference, initiated by ELRA, are progressing according to plans. The conference will focus on the following issues: the availability of language resources and the methods for the evaluation of resources, technologies and products, for written and spoken language. Substantial mutual benefits can be expected from addressing issues like these through international co-operation. The aim of this Conference is to provide an overview of the state-of-the-art, discuss problems and opportunities, exchange information regarding ongoing and planned activities, language resources and their applications, discuss evaluation methodologies and demonstrate evaluation tools, explore possibilities and promote initiatives for international co-operation. As of today, late February, more than 200 papers have been accepted by the Program Committee, and they will be presented in oral, poster or demo sessions at the conference. The programme will be presented in March and published on the ELRA (http://www.icp.grenet.fr/ELRA/home.html) and LREC Web-sites (http://ceres.ugr.es/~rubio/elra.html).

In parallel with the conference, an exhibition will be organised by ELRA. This exhibition is open to companies and projects wishing to promote, present and demonstrate their products and prototypes to the wide range of experts and representatives from all over the world participating in the conference. For more information on this, please contact the ELDA office on elra-elda@calva.net. As for activities at the conference itself, the following panels, pre- and post-conference workshops will be held. For complete information on the workshops, please consult http://www.icp.grenet.fr/ELRA/home.html

Panel of the Funding Agencies: Members of the major agencies funding research and development in Language Engineering (NSF, ARPA, EC, etc.) will discuss priorities and perspectives for international co-operation. *Lexical Semantic Standards for Information Systems:* The panel will discuss guidelines for the standardisation of lexical encoding with specific reference to requirements for Machine Translation and Information Systems. *Industrial and R&D use of Language Resources:* Users and providers of Language Resources, from industrial companies and from the public research sector, will discuss the priorities and the economical aspects of producing, distributing and using Language Resources, and the importance of their availability.

Linguistic Coreference

26 May 1998, morning session

Aims: It is essential, for a natural language processing system, to instantiate each object, process, attribute, and property correctly, so that all references to the same item be recognised as such and an inventory of all distinct items be accurate at all times. This problem is far from being resolved. There are both linguistic and computational reasons for this deficiency. First, there is no satisfactory microtheory of linguistic coreference. Secondly, there is no satisfactory application of such a microtheory to NLP. One persistent problem throughout the existing computational ventures into coreference has been the lack of a consistent theoretical approach to it. What is needed for a full, accurate, and reliable approach to coreference can be summarised, somewhat schematically, as involving the following steps: 1. understanding fully the range of the phenomenon and of the rules that govern it (theory); 2. determining the extent of machine-tractable information in the rules; 3. taking stock of all the rules that can be computed; 4. developing the appropriate heuristics for the computable rules; 5. computing the rules.

Contact: Dr. Sara J Shelton at:
sjshelt@afterlife.ncsc.mil

Adapting Lexical and Corpus Resources to Sublanguages and Applications

26 May 1998, morning session

The workshop will provide a forum for those researchers involved in the development of methods to integrate corpora and MRDs, with the aim of adding adaptive capabilities to existing linguistic resources. The Central topics are: corpus-driven tuning of MRDs to optimise domain-specific inferences; terminology and jargon acquisition; sense extensions; acquisition of preference or subcategorization information from corpora; taxonomy adaptation; statistical weighting of senses etc. to domains; use of MRDs to provide explanations of linguistic phenomena in corpora; what is the scope of "lexical tuning"; the evaluation of lexical tuning as a separate task, or as part of a more generic task.

Contact: Prof. Paola Velardi at :
velardi@dsi.uniroma1.it

Minimising the Effort for Language Resource Acquisition

26 May 1998, afternoon session

The proposed workshop will be devoted to ANY TECHNOLOGICAL AND ADMINISTRATIVE FACET OF ECONOMY OF ACQUISITION EFFORT. Examples are: minimisation of effort in acquiring monolingual and multilingual text corpora; minimisation of effort in acquiring computational lexicons; minimisation of effort in acquisition of resources for the support of corpus-based language engineering methods; minimisation of effort in acquiring grammatical coverage of languages and sublanguages; methods of determining levels of reusability of existing language resources; balancing the needs of the application and the grain size of language description; minimisation of effort through balancing automatic and interactive methods of knowledge acquisition; evaluation of potential utility of resources to applications.

Contact: Svetlana Sheremetyeva at :
lana@crl.nmsu.edu

The Evaluation of Parsing Systems

26 May 1998, afternoon session

This workshop will provide a forum for researchers interested in the development and evaluation of natural language grammars and parsing systems, and in the creation of syntactically annotated reference corpora. The aim of this workshop is to provide a forum for discussion of evaluation methods for parsing systems, and proposals for the development of syntactically annotated language resources. The topics are: descriptions of generic syntactic annotation schemes; methodologies and metrics for parsing system evaluation; reports and analyses of the results of utilising particular parser evaluation schemes; description/analysis/experience of language-dependent and task-dependent syntactic annotation schemes.

Contact: John Carroll at :
john.carroll@cogs.susx.ac.uk

Towards a European Evaluation Infrastructure for NL and Speech

27 May 1998, morning session

Right now, a generic framework for semi-automatic quantitative black-box evaluation of Speech and NLP systems does not exist in Europe. When confronted to a choice,

developers and users prefer to ask the opinion of local experts as any other way of processing is either unrealistic or too costly. The LE4 project ELSE aims at providing developers with a generic strategy and definition of the primary building blocks needed to implement a semi-automatic quantitative black-box evaluation scheme. Topics include the multilingual nature of evaluation, lessons from the past, the need for language resources.

Contact: Steven Krauwer at:
steven.krauwer@let.ruu.nl

Language Resources for European Minority Languages

27 May 1998, morning session

The minority or "lesser used" languages of Europe (e.g. Basque, Welsh, Breton) are under increasing pressure from the major languages. Some of them (e.g. Gaelic) are becoming endangered, but others (e.g. Catalan) are in a stronger position, with a certain amount of official recognition and funding. Some minority languages have been adequately researched linguistically, but most have not, and the vast majority do not yet possess basic speech and language resources which are sufficient to permit commercial development of products. The aim is to make it easier for isolated researchers with little funding and no existing corpora to begin developing a usable speech or text database.

Contact: Briony Williams at:
briony@cstr.ed.ac.uk.

Speech Database Development for Central and Eastern European Languages

27 May 1998, afternoon session

Speech databases have been produced for a number of the world's major languages, but most languages of Central and Eastern Europe have received little attention in international terms until recently, though they are of major importance for the future of European speech science. The topics are: Recording techniques and standards; Available software tools; Annotation, transcription and labelling; Automated time-alignment of labels; Phonetic problems of languages of Central and Eastern Europe; Quality control; Requirements for larger-scale databases; Dissemination of data, recording further languages, possibilities for future collaboration.

Contact: Peter Roach at:
p.j.roach@reading.ac.uk

Distributing and Accessing Linguistic Resources

27 May 1998, afternoon session

This workshop will discuss ways to increase the efficacy of linguistic resource distribution and programmatic access, and work towards the definition of a new method for these tasks based on distributed processing and object-oriented modelling with deployment on the WWW. In general the reuse of NLP data resources has exceeded that of algorithmic resources. However, there are still two barriers to data resource reuse which will be discussed: 1) each resource has its own representation syntax and corresponding programmatic access mode; 2) resources must generally be installed locally to be usable. This workshop will discuss ways to overcome these barriers.

Contact: Yorick Wilks at:
yorick@dcs.shef.ac.uk

Translingual Information Management: Current Levels and Future Abilities

30 May - 1 June 1998

This is a two-day post conference workshop to be held after the LREC.

Contact: Nancy Ide at:
ide@cs.vassar.edu

Practical Automatic Dictation Systems

Melvyn J. Hunt

Text creation is possibly the most frequent purpose of human interaction with computers. It is therefore natural to want to simplify this process by providing automatic transcription of spoken input. A generation ago, it was common to have a professional typist transcribe an audio recording or a longhand manuscript produced by the originator of the text. This practice has now all but disappeared, putting pressure on those who must create documents but are not skilled typists, and consequently raising the interest in automatic dictation.

The Emergence of Commercial General-Purpose Automatic Dictation Systems

Although there were some earlier systems for composing text in narrowly defined areas, such as radiology reports, the first general-purpose commercial dictation system (DragonDictate®) went on sale in 1990. British English and German versions followed the original American English offering. By today's standards, they were expensive and required the installation of a special audio processing board containing a digital signal processor. In addition, the user had to invest significant effort to adapt the system to his or her voice before recognition accuracy reached a useful level, which was still relatively low by today's standards. For these reasons, a high proportion of early users were people who had a strong motivation, because they did not have the option of using the keyboard, either because of paralysis, or because arthritis or RSI (Repretitive Strain Injust) made typing painful. For many such users, the first appearance of a hands-free method of creating text made a return to productive work possible.

The success of this product was helped by the emergence of a dominant PC architecture, the IBM-compatible PC, and a dominant operating system, Microsoft DOS. Similarly, the next major advance, the elimination of the need for special hardware, was enabled by the emergence of a standard for multimedia audio input, the widespread inclusion of audio input in PCs, and the faster Intel 486 processor. The first such "software-only" system, DragonDictate for Windows, appeared in 1994. As well as offering text creation, it allowed hands-free control of virtually any Windows software, being capable of recognising the vocabulary appearing in menus and dialogue boxes. Software-only dictation systems for Microsoft Windows® were subsequently produced by Kurzweil and by IBM, who had previously offered dictation systems for their OS/2™ operating system and RS/6000™ Unix™ workstations.

By 1996, Windows-based general-purpose dictation systems were available from three sources, in a variety of languages including American and British English, French, German, Spanish, Italian, Swedish and Arabic. In addition, a dictation system for the Apple Mac was available from Articulate Systems. Although laboratory systems for continuous dictation had been demonstrated for some time, and although some of the commercial general-purpose dictation systems allowed continuous input of commands and number strings, they still required a pause between each word during general text creation, albeit of shorter duration than was the case in the earliest products.

In addition, Philips had launched a topic-specific true continuous dictation system for use

in hospitals. It differed fundamentally from the systems described above in that its existence was effectively invisible to the person creating the document. He or she simply continued to make audio recordings for transcription, but the transcription process was accelerated by having a topic-specific continuous speech recognition system produce a first draft, which was then checked and corrected by a human audio typist.

The Philips system, at least its initial version, needed special processing hardware (ASICs). In early 1997, however, IBM launched a topic-specific continuous speech recogniser that ran on a PC without requiring special hardware, and a similar product using a Dragon speech recognition engine followed.

The first general-purpose continuous dictation system was launched in April 1997 and began shipping in June of that year. This was Dragon NaturallySpeaking™, and IBM's general-purpose continuous dictation system, ViaVoice™, followed a few months later. Both worked on PCs, albeit PCs at the high end of what was around at the time of their launches, and both worked with Windows 95™ and with Windows NT™. By the end of last year, general-purpose continuous speech recognition systems were available from both Dragon and IBM in most major Western European languages, as well as in American English, and Kurzweil, now owned by Lernout and Hauspie, had announced their own product, which they called Voice Xpress™.

One of the most recent entrants into the general-purpose automatic dictation market is a company called Speech Machines, which acknowledges the British Government's Speech Research Unit in Malvern and Cambridge University Engineering Department as sources of its speech recognition technology. Like the earlier Philips product, Speech Machines offer an off-line non-interactive approach, in which the automatically produced first draft is verified and corrected by audio typists. It differs, however, in that it is offered as a bureau service, rather than as a product in itself, and in that it accepts telephone speech, with the completed document being returned via the Internet.

The Technology

Although manufacturers do not always disclose details of the technical basis of their products, it is reasonable to infer that most, if not all, commercial dictation systems follow a similar basic technical approach.

The statistical technique of Hidden Markov Modelling is used in the central process of recognising the phonetic identity of the speech sounds to be recognised. The fundamental reference units against which the speech is compared represent phonemes in particular phonemic contexts. These units are built up into words using a lexicon providing a phonemic transcription of the vocabulary of the recogniser. The comparison between the spoken input and any hypothesised word or word sequence can then be given an acoustic match score. Typically, continuous and isolated-word dictation systems have active vocabularies ranging in size from 20,000 to 60,000 words, with additional backup vocabularies, whose contents can be accessed during correction, comprising up to around 200,000 words.

Individual voices clearly differ because of regional accents, physiology and simply idiosyncratic differences. For optimum recognition performance, such differences have to be taken into account. The original phonetic models supplied with a dictation system are normally derived from a broad population of male and female speakers. These models are then adapted to the user. Data for adaptation can be derived either from an explicit training session, in which the user is prompted to read aloud a story or some other text, or implicitly during use. In principle, every word spoken to the dictation system can be used for adaptation, with the assumption that if a word recognised is not corrected, then it must have been recognised correctly. This is the fastest way to gather adaptation material during use. However, if the user fails to correct recognition errors quickly, false adaptation can occur, potentially leading to a degradation in performance. For this reason, systems now often allow adaptation during use to occur only after explicit correction.

Even after adaptation of the phonetic models to the user, identification of words with an acceptable recognition accuracy in such large vocabularies remains a challenging task. The evidence from their acoustic properties has to be augmented by a so-called language model that reflects the fact that the sequence of words being dictated is not, in general, arbitrary. However, in contrast to some command-and-control applications of speech recognition, dictation systems must ultimately be prepared to accept any sequence of words, even unlikely and ungrammatical sequences. Consequently, rather than being based on unbreakable rules, the language model for general-purpose dictation systems is necessarily statistical, providing an estimate of the probability of any given sequence of words. The resulting language model score for an interpretation of some spoken input can then be combined with the acoustic match scores to determine the most probable interpretation of an utterance. In interactive dictation systems, alternative interpretations ordered by their combined match scores can be made available to the user, facilitating correction of recognition errors.

The narrower the range of language used in the dictated text, the tighter the language model can be, and the easier the recognition task becomes. This is why general-purpose dictation systems address a much more challenging task than those offering topic-specific text creation. Typically, interactive dictation systems adapt their language model to the habits of the user. Some of the continuous-speech systems can take existing documents and use them to adapt the language model even before the first text has been produced. Dragon NaturallySpeaking, for example, is able to scan sets of documents, updating its statistical data and ensuring that any words encountered are present in the active vocabulary, either by moving them from the backup vocabulary, or, if they are completely new, guessing their pronunciations and optionally allowing the user to provide a spoken example. One recent version allows the user to develop multiple language models. There can be one, for example, for producing reports in a specialised professional field, a second for general domestic correspondence, and perhaps a third for, say, producing the minutes of meetings of the local branch of a charitable organisation.

Because the language model for dictation systems can never rule out the occurrence of any word, but merely bias the decision process against accepting it, the whole active vocabulary must always be considered, potentially creating a very heavy computational load. The problem gets worse with continuous speech recognisers, where words can start anywhere in a speech stream, not just after a pause. Developers of dictation systems have solved this problem by introducing "rapid match" techniques that quickly reduce the set of possible words to a manageable short-list for detailed consideration. These techniques are the key to practical performance on affordable hardware, and some of them have been patented.

Language Differences

The English-speaking world has no monopoly on advances in speech recognition. Nevertheless, vastly more work has been done on recognition of English speech than on any other single language. Most commercial speech recognition systems have appeared in English first. It is possible that as a

result the development of the technology has been biased towards linguistic features present in English. For example, most large-vocabulary recognition systems that have been described treat each inflected variant of a word as an independent entity in the lexicon. This is perfectly practical for English, where we have two or three forms of each noun, typically four forms for each verb, and only one form of each adjective. However, in languages such as Finnish, Hungarian, Estonian and Turkish, with their multiplicity of inflected forms, treating each variant as an independent entity would lead to an immense increase in active vocabulary size. Moreover, with many highly inflected languages, including some Indo-European languages such as Russian, word order within the sentence is not constrained by grammar, grammatical function being indicated by the inflection. This poses a problem for the current syntactically based language modelling.

However, even if dictation systems have been developed for English initially, it is not the case that they always work better in English than in any other language. Indeed, several tests have found recognition accuracy of commercial dictation systems to be higher in Italian than in English, presumably because of the simpler syllabic structure, and the absence in Italian of the tendency in English to centralise and even suppress vowels in unstressed syllables.

At the other end of the scale of Western European languages for which commercial dictation systems are available lies French. The difficulty in obtaining accurate recognition of French stems largely from the very high proportion of homophones in the language, some spanning semantic differences (e.g. *ver*, *verre*, *vert*, *vers* — "worm", "glass", "green", "towards"), others grammatical differences (e.g. singular/plural differences in nouns and adjectives, many verb inflections and past participle agreements). The latter can extend over a long range (e.g. *les maisons de pierre rouge qu'ils n'avaient jamais vues*), posing a serious challenge to our language models, which tend to rely on short-range syntactic behaviour.

German is generally found to lie between French and English in difficulty. Word compounding, especially in nouns, is probably the biggest challenge in this language. The dictation system must either require the user to indicate the beginning and ending of a compound word, or try to work out for itself whether a sequence of words should be returned as a single unit or not. German also has a moderate degree of inflection of nouns, adjectives, pronouns and articles, and the distinction between the unstressed *en* and *em* endings is not acoustically very salient.

Although British and American varieties of English are variants of a single language by most definitions, they are often treated as separate languages for automatic dictation systems. Pronunciations differ both systematically and in unpredictable ways for some specific words (e.g. *schedule*, *tomato*, *vase*...). Some vocabulary items are different (e.g. *aeroplane/airplane*, *zed/zee*...) and hundreds differ in their spelling. The names of some punctuation marks (e.g., *full stop/period*, *bracket/parenthesis*) are also different. Finally, the frequencies of words such as place names and names for national institutions are quite different in the two varieties of the language. Varieties of English spoken in other parts of the world are not normally treated separately at present.

Before concluding this section, it may be worth pointing out that automatic dictation has a particular advantage over keyboard input for people working in more than one language, since keyboard layout is often different between different language communities, and the sets of diacritics (e.g. umlauts, tildes, cedillas and accents) and special letter symbols that are peculiar to a particular Western European language tend not to be easy to produce on keyboards designed for another language. Spoken input, of course, is not subject to problems in this area.

Practical Characteristics of Current Automatic Dictation Systems

Since the overwhelming proportion of automatic dictation systems in use are interactive, rather than the off-line kind, the discussion from now on will be confined to interactive systems. Such systems normally do more than just allow the user to transmit a sequence of words to the screen. They may allow a document to be corrected, edited and formatted by voice, and features of software applications to be controlled by voice. Voice macros may allow frequently used blocks of text to be produced in response to a single brief spoken command. Some can include complex formatting operations, or even operations such as sending a fax to a named addressee.

With some recent systems, the speech that has been input to the dictation system can be played back, and a spoken version of the text that has been composed can be read aloud using a text-to-speech synthesis system. These features have obvious advantages for those with vision difficulties and those who prefer not to look at the screen. They are also useful, however, for the general user. Playback of input speech allows correction of recognition errors after a long delay when the user may have forgotten what was actually said. Reading back the text with synthesised speech can be useful for proof-reading. An important class of recognition errors with continuous speech dictation systems consists of substitutions or deletions of small, common words. Unlike typing errors, errors made by continuous dictation systems are inevitably correctly spelled and plausible in their immediate context, making them sometimes hard to spot by eye. When read aloud, however, they are much more evident.

Current dictation systems continue to use headset-mounted microphones connected by wire to the computer. Although hand-held and stand-mounted microphones can be used successfully (and Philips have developed a hand-held microphone specifically for computer input for dictation applications), the headset mounts continue to be preferred, because they allow the use of pressure-gradient microphones. These microphones are sensitive to local sources of sound and much less sensitive to distant sources, making it possible to use them in environments containing noise and other speech, without spurious recognition occurring.

Public Acceptance of Automatic Dictation

The very fact that speech is such a natural and effortless mode of communication between people sometimes erects a barrier to its public acceptance in creating text. We accept without difficulty that it is worth the effort to learn the much less natural practice of pressing down little plastic pegs with our fingers in order to create text, but having to learn to speak clearly, pausing between each word, often seems an unreasonable requirement. Moreover, we are ourselves such brilliant decoders of the speech waveform when it corresponds to a grammatical and meaningful sequence of words, that we often have unreasonably high expectations of the performance of automatic systems and find any errors that they make to be unreasonable.

For most of those who are not expert typists, even isolated-word automatic dictation systems probably represent a faster and less tiring method of creating text than typing, but they do require some initial commitment. After a three-month trial of DragonDictate in several languages at the European Commission Translation Service, for example, no fewer than 12 of the 14 subjects in the trial said that they intended to continue using the isolated-word dictation system. However, these subjects were unusual in

having contact with each other, as well as the support and interest of their employer. Unlike someone who begins to use a well-known word processing package, who will be surrounded by others who are already successfully using it and can provide help, those who have tried using automatic dictation systems in the past have often been isolated pioneers in their organisations. In these circumstances, when difficulties were encountered, it was all too easy to abandon the attempt and return to two-finger typing.

Since the appearance of general-purpose continuous-speech dictation systems last year, there is evidence that the situation is changing dramatically. The speaking style is much more natural, removing one of the major barriers to acceptance. The systems are also much faster. In public contests between Dragon NaturallySpeaking and expert professional typists in both the US and in Britain, no typist has ever won. Reviews in the media, particularly in the US, have been enthusiastic. Perhaps more significantly, many print journalists said that they were producing their review using the very product that they were reviewing!

A recent market survey carried out for Dragon Systems in the US found that the majority of users of NaturallySpeaking worked in medicine, law, education, or business, and that less than 10% had a physical disability or feared developing one through typing. Over 90% of users said that they would recommend the system to others.

In that same survey, over half the users said that they had bought new hardware, either a new PC or additional memory, in order to run the software. This suggests that bundling dictation software with PCs may well influence buying decisions, and indeed bundling isolated-word or continuous-speech software with PCs does appear to be becoming increasingly common.

Until last June, Dragon Systems had not sold its products through retail channels in the US. Just six months later, the value of its monthly sales ranked number 13 in the list of all companies selling retail business software products of any kind in the US.

It looks as though the chain reaction needed for the general acceptance of a new mode of communication with PCs may have started.

Future Prospects

This article has confined its attention to dictation, but we should not forget that there will be an increasing number of exciting applications of speech recognition in areas other than dictation.

Within the dictation application area, over the next few years we will undoubtedly see the increasing development of remote and distributed systems, and perhaps of complete dictation systems in palmtop computers, where efficient keyboard input is simply not an option.

Perhaps we will see the transcription of speech not primarily intended for dictation, such as the transcription of court and parliamentary proceedings, and as aids for the hearing-impaired.

For some people, the current microphone arrangement continues to be a barrier to acceptance. We will probably see the widespread adoption of wireless microphones and possibly desk-mounted microphone tracking arrays, which can offer some of the advantages of a pressure-gradient microphone without the inconvenience of having to wear a headset.

Even if there are no further technical developments, though, the chain reaction needed for widespread acceptance may already be unstoppable.

Melvyn J. Hunt
Dragon Systems UK
E-mail: melvyn@dragonsys.com

EuroWordNet: Building a Multilingual Database with Wordnets for European Languages, *Piek Vossen*

All the knowledge and information in the Information Society is useless unless we are able to communicate with the keepers of it: computer systems. Most of the information they hold is stored as text and pictures which people may understand, but computers do not. It is clear that morphosyntactic analysis and speech processing will not get us very far in exploiting this information. Statistical techniques have been more successful, especially in information retrieval, mainly because they are computationally tractable, do not rely on expensive resources and can be applied to any domain that contains large quantities of text. Nevertheless, the benefits of shallow statistical processing are limited, and the time seems ripe for exploring a more content-driven method for processing information.

It is only fair to say that the area of semantics and interpretation includes many hurdles and pitfalls that make it difficult to define its limits and scope. Meaning is said to be fuzzy, complex, context-dependent, knowledge-dependent, and ambiguous. Still, some recent projects, such as the development of WordNet, EDR, MikroKosmos and Cyc, have shown that it is possible to develop feasible large-scale resources involving at least some of the required knowledge. These resources are being used, showing that it is not necessary to know the full scope of the problem to do useful things. What is more, we will only be able to tackle the full problem when we start dealing with parts of it in a realistic applied environment.

In Europe, these resources are not (yet) available in most languages. An additional problem is multilinguality. The European Information Society not only needs these resources in every language, it also needs mapping across every language resource. This is an absolute prerequisite for its successful development. EuroWordNet directly addresses this problem by developing a multilingual database with wordnets for a large set of European languages. Each of these wordnets is structured along the same lines as the Princeton WordNet, i.e. around the notion of a synset. A synset is a set of synonymous word meanings between which basic semantic relations are expressed, for example, hyponymy (car – vehicle), meronymy (wheeled vehicle – wheel) and cause (kill – die). In addition to the relations between synsets, the so-called language-internal relations, each synset in EuroWordNet is also linked to the Inter-Lingual Index or ILI, thus constituting a multilingual database (see Figure 1.). This ILI is an unstructured list of concepts, called ILI records, mainly taken from WordNet1.5, but adapted to improve the matching of synsets across languages. Although the ILI as such will not be structured in terms of semantic relations between the concepts, it will nevertheless give access to a shared top-ontology and a domain-ontology. These ontologies are applied to particular sets of ILI records, and, in principle, apply to any language-specific synset that is related to these ILI records.

Using the ILI, it is possible to go from a synset in one wordnet to the synsets in the other wordnets that are related to the same ILI record, and to compare the lexical semantic structures. A comparison of a large set of wordnets will give an indication of the differences in the relations across the wordnets. These differences can either be inconsistencies, or they can point to language-specific differences in the resources. The fact that we link a whole series of wordnets to the ILI

makes it possible to develop a more fundamental view on these differences, helping to understand how language-specific the wordnets are and pointing to areas where work remains to be done. The proportion of lexical semantic relations that is shared by a large number of wordnets gives a good indication about the quality of the relations. Special interfaces have been developed in the EuroWordNet database to carry out this kind of comparison.

The first project consortium (LE2-4003) has worked on the Dutch, Italian and Spanish wordnets, while the English wordnet was only adapted for relations which were not covered in the Princeton WordNet1.5. Recently, the project has been extended (LE4-8328) to include French, German, Czech and Estonian. The wordnets are built from existing resources as far as possible, covering the general, generic vocabulary of the languages. The languages in the first project (LE2-4003) aim at a size of 30,000 synsets and 50,000 word senses. The languages in the extension will aim at a set of 15,000 synsets and 30,000 word meanings. Finally, the wordnets will be validated by three users in (cross-linguistic) Information Retrieval (IR) applications. The validation tools as such will not be developed; instead, the wordnets will be loaded into existing IR systems. Further information on the project and the participants can be found at the EuroWordNet Website (http://www.let.uva.nl/~ewn).

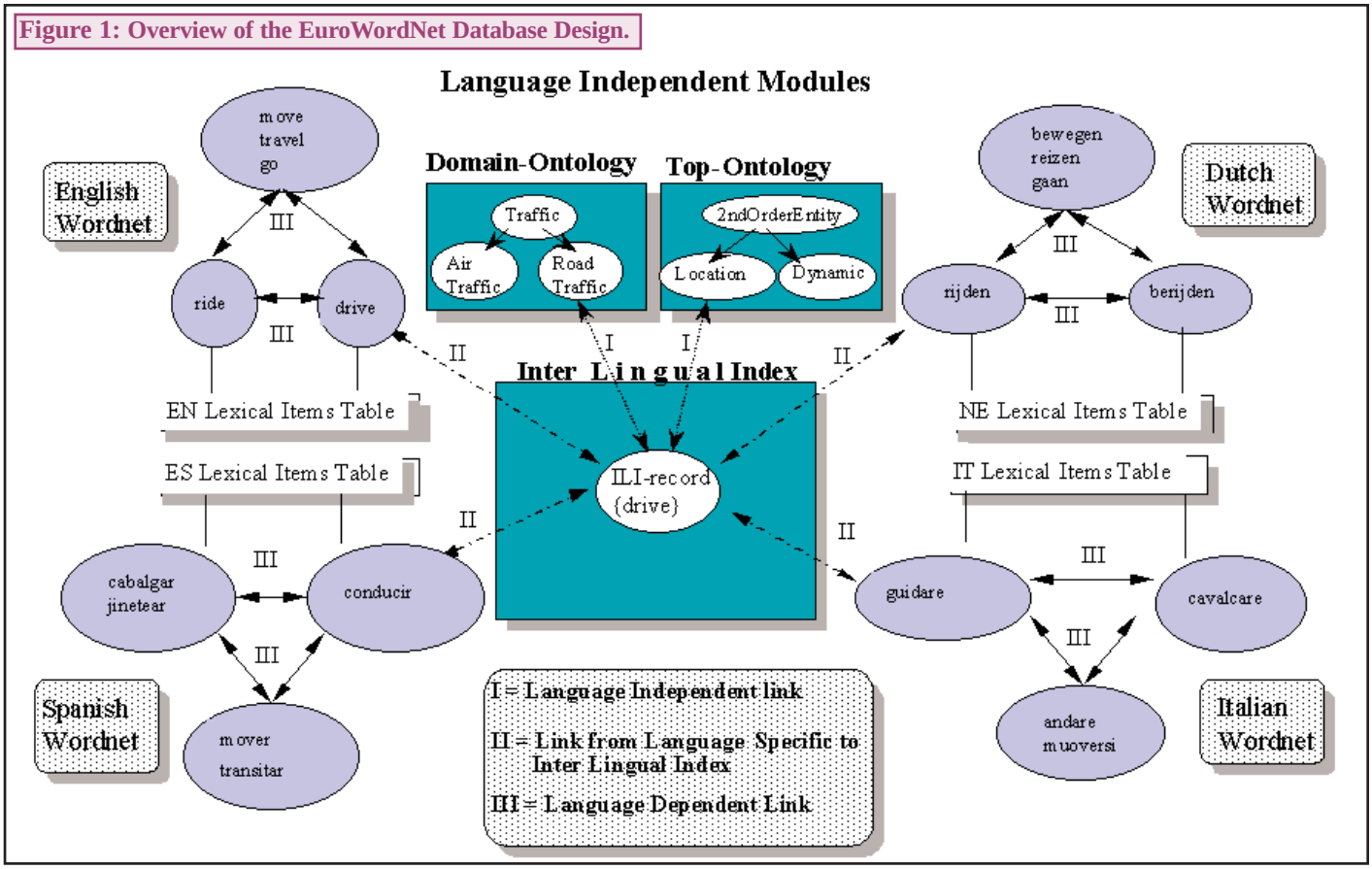
Wordnets as autonomous language-specific networks

An important characteristics of the project is that the wordnets are treated as autonomous systems of language-internal relations. This will give us the flexibility to develop the wordnets relatively independently; this is necessary because each group has a different starting point in terms of resources, tools and databases. However, there is also a more fundamental reason why we take this position. Each wordnet represents a unique network of relations, due to the lexicalisation patterns that are specific to the languages. For example, in the Dutch wordnet we see that *hond* (dog) is classified as both *huisdier* (pet) and *zoogdier* (mammal). However, there is no equivalent for *pet* in Italian, and likewise the Italian *cane*, which is linked to the same synset *dog*, is only classified as a *mammal* in the Italian wordnet. In EuroWordNet, we take the position that it must be possible to reflect such differences in lexical semantic relations. The wordnets are seen as linguistic ontologies rather than ontologies for making inferences only. In an inference-based ontology it may be the case that a particular level or structuring is required to achieve better control or performance, or a more compact and coherent structure. For this purpose it may be necessary to introduce artificial levels for concepts which are not lexicalised in a language (e.g. *natural object*, *external body parts*), or it may be necessary to neglect levels which are lexicalised, but not relevant for the purpose of the ontology. A linguistic ontology, on the other hand, exactly reflects the lexicalisation and the relations between the words in a language. It is a "wordnet" in the true sense of the word, and therefore captures valuable information about the expressiveness of languages: the words and expressions available in a language.

The difference is illustrated in Figure 2, where the hyponymic structure of WordNet1.5 reflects a combination of lexicalised and non-lexicalised categories and the Dutch Wordnet only contains categories lexicalised in the language. In WordNet1.5 we see that the synset for *object* is first subdivided into two subclasses, *artifact* and *natural object*, of which the latter is not a lexicalised expression in English (i.e. an expression you would expect to find in a dictionary), but rather a regularly composed expression. The class *artifact* has an important subclass (*instrumentality*) which is used to group related synsets such as *implement*, *device*, *tool* and *instrument* under a common denominator. Such a grouping seems helpful in organising the hierarchy and predicting the functionality of the subclasses. However, it does not give correct predictions about the substitutability of the nouns: you cannot refer to *containers*, *boxes*, *spoons*, and *bags* using the noun *instrumentality* in English.

In the Dutch hierarchy, we see that artificial levels such as *natural object* and *instrumen-*

Figure 1: Overview of the EuroWordNet Database Design.



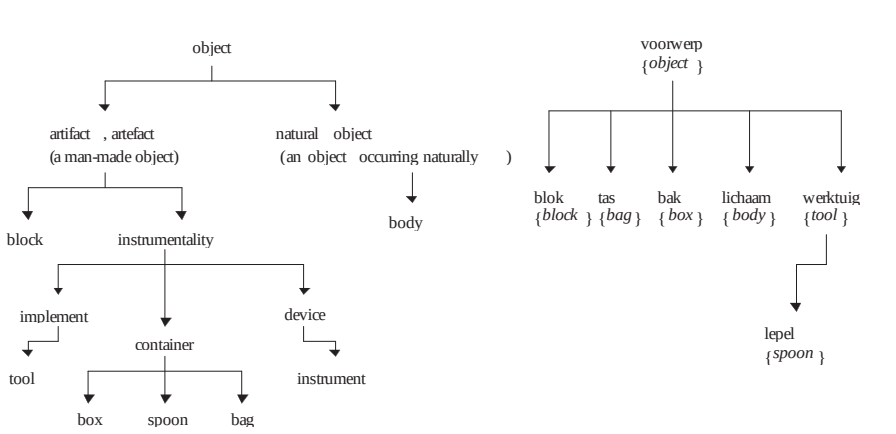


Figure 2: Lexicalised and Non-lexicalised Levels in Wordnets.

tality have not been used. Furthermore, there are no exact equivalents for *artifact* and *container* in Dutch (The word *container* does exist in Dutch but is only used for big containers on ships or for big garbage cans). As a result of this, we get a much flatter hierarchy from which particular properties such as *natural*, *artificial* and *functionality* cannot be derived. On the other hand, the network correctly predicts the expressive capacity of Dutch, because it only includes the legitimate words (and expressions) of the language. We could invent new classes and expressions in Dutch to capture different generalisations, we could even take over the WordNet1.5 classes, but there are no a priori criteria to decide what are useful classes and what are not. We may end up adding any conceivable semantic property as a class to create very rich inheritance structures, or we may take over all possible classifications from all the other wordnets. However, this would destroy the wordnet as a network of legitimate expressions in a language, and it would still not automatically give us a good conceptual ontology for inheriting properties.

In addition, it is possible to extend the database with a separate language-neutral ontology which takes care of the inferences and is well designed for that purpose. When this ontology is linked to the ILI, all the wordnets can access the classifications there to find the correct inferences for the synsets. The wordnets then provide the precise mapping of the language-specific vocabulary on this ontology. To get at such ontology, we are co-operating with the ANSI Group on Standardising Ontologies, which is developing a standardised Reference Ontology.

The top-down construction of the wordnets

A drawback of the flexible design described above is that the interpretation and coverage of the wordnets may easily drift apart. There is no guarantee that the same conceptual areas are covered or that the relations are encoded in the same way. To minimise this danger, the wordnets are developed top-down starting with a shared set of Base Concepts. These Base Concepts have been selected for their importance in the local wordnets. Importance has been measured in terms of the number of relations and the position in the hierarchy. The more relations or the higher the position, the more important a meaning is. All meanings which play a major role in at least two wordnets have been selected. This has resulted in a set of 1,059 Base Concepts, represented as WordNet1.5 synsets. The Base Concepts have been described using a top-ontology with 63 basic semantic distinctions (Top Concepts) such as *Substance*, *Object*, *Artifact*, *Natural*, *Function*, *Dynamic*, *Static*, *Cause*, *Location*, *Experience*. The top-ontology has been based on other available ontologies and has been adapted to reflect the diversity of the Base Concept selection. The classification of the Base Concepts in terms of the Top-Ontology provides a common framework for the development of the individual wordnets by the different sites.

The actual building of the separate wordnets then takes place as follows:

1. Selection of a well defined set of word meanings. Encoding of lexical semantic relations and equivalence relations for this set.

2. Conversion of the data to the EuroWordNet import format.

3. Loading the data in the EuroWordNet database.

4. Comparison of the wordnets for particular subsets.

5. Revision of the wordnets in the EuroWordNet database.

6. Extension of the first selection.

First, each group has determined the synsets that most closely represent the common Base Concepts in their local language, given the available resources. This selection has been extended with other meanings which are important in the local wordnets, but which are not part of the common set of Base Concepts. This set of meanings has been classified in the local wordnets in terms of their hyperonyms, resulting in a unified tree. Note that these classifications may be different from wordnet to wordnet and still be compatible with the top-ontology classification. In addition to this top layer, we have included those hyponyms that are also (important) hyperonyms of more specific meanings. Together this selection represents the core of each wordnet with the most important meanings on which the remainder of the vocabulary depends. To summarise, each core wordnet includes at least:

1. The best representatives for the 1,059 Base Concepts.
2. Other meanings important for the local wordnet.
3. Hyperonyms for the local Base Concepts.
4. Most important hyponyms of the local Base Concepts.

The core wordnets are specified at least for synonymy, hyponymy and their equivalence relation to the ILI. Optionally, any other salient relation has been encoded to *interconnect* the meanings in the wordnet. Because of the importance for the total wordnets, the manual work has been focused on these cores. The extension from the core wordnets will be done top-down using semi-automatic techniques. Currently, the top-ontology, the Base Concepts and the core wordnets have been finalised for Dutch, Italian and Spanish. The data have been loaded in the EuroWordNet database and are being compared. From the comparison in the EuroWordNet database, it may follow that particular relations or word meanings are missing, that they have to be revised, or that equivalence relations are not correct. This will lead to a modification of the core wordnets. In the remainder of the project, the cores will be extended, and the other languages will be added. The new languages will first develop similar core wordnets and extend them in a later phase.

A separate task is the adaptation of the ILI. In practice it turns out to be difficult to find a precise matching between a synset in the local wordnet and a synset in the ILI (mostly synsets taken from WordNet1.5). The Base Concepts, which are often vague and polysemous, are difficult to match. In many cases there will be a many-to-many matching, or there will not be an equivalent concept in the ILI. To improve the matching, the ILI is then adapted. There are two types of modification:

1. Adding of new concepts which are missing.
2. Creation of sense groups between closely related senses or regular polysemy.

The addition of new concepts is necessary to enable a precise mapping of synsets across

wordnets in cases that there is no such concept in WordNet1.5. For example, if only the Spanish and Italian wordnet include a meaning for a type of *wine*, the new concepts should make it possible to specify the equivalence between Spanish and Italian, despite the absence in English.

The sense groups are necessary to deal with inconsistent and fuzzy sense differentiation across the lexical resources. We often see that resources only specify one out of several meanings that can be distinguished (often on a regular basis), e.g. "embassy" as an *institute* or as a *building*. This may mean that concepts cannot be linked across languages because different meanings are represented, i.e. either the *institute* or the *building*. To relate these meanings across the wordnets, we must extend the ILI with a globalized sense in which these meanings are grouped: "embassy", both as a *building* and an *institute*. Each synset in the local wordnet linked to one of the more specific meanings will then get an additional equivalence link to the globalized meaning. These equivalence relations are differentiated from the normal equivalence relations so that it is possible to use these more global matches if a more precise matching gives no result.

Availability

The shared components, such as the top-ontology, the ILI and the selection of Base Concepts, will be freely available. The same holds for the multi-lingual viewer that can be used to access the database version of these components and the wordnets. The language-specific wordnets will be property of the builders, in some cases in combination with the providers of the background resources. All components will be available both in database format and as ASCII files. All wordnets can be licensed either from ELRA or from the owners, and the tool for building and extending the databases can be licensed separately. The core wordnets will be available from ELRA; we expect this for the beginning of 1999.

Piek Vossen
Universiteit van Amsterdam
Spuistraat 134, 1012 VBAmsterdam
The Netherlands
Tel: +31 20 525 4669 - Fax: +31 20 525 4429
E-mail: piek.vossen@let.uva.nl

Multilingual Natural Language Processing at XRCE

Frédérique Segond, Gregory Grefenstette and Annie Zaenen

Xerox Research Centre Europe (XRCE) pursues a vision of document technology in which neither physical location, nor language nor medium - electronic, paper, or other - imposes a barrier to effective use. Our primary activity is research. Our second activity is a program of advanced technology development, to create new document services based on our own research. The linguistic technologies are then commercialised through InXight (LinguistX) and Xtras. On top of this, Xerox has just established a new entity to coordinate the development, maintenance and internal distribution of lingware for the whole corporation. This entity is based in Grenoble, France, and will primarily concentrate on multilingual resources such as morphological analysers, bilingual dictionaries and part-of-speech taggers. The entity will facilitate the sharing of language resources among various Xerox groups, be they involved in research, development, products, or services. We also participate actively in exchange programs with European partners. Language issues represent important aspects of the production and use of documents. As such, language is a central theme of our research activities.

The Multilingual Theory and Technology (MLTT) team is dedicated to the study of natural language processing for different languages. At the moment we have developed tools for more than 10 languages.

This article gives an overview of our work in developing multilingual language processing tools: tokenisers, morphological analysers, part-of-speech taggers and incremental finite-state parsers. We also briefly describe a number of research applications (corpus-based terminology extraction, comprehension and translation aids, and cross-language information retrieval) derived from the use of these tools.

Multilingual language processing tools

Our natural language processing tools integrate both statistics and linguistic knowledge. We believe that improvements in the field of NLP will be made by combining the two methodologies in the best possible way.

The finite-state approach has theoretical and practical advantages. The theoretical advantage is that finite-state machines are well understood mathematical entities, with well-known properties. Finite-state transducers can be composed, intersected, or unioned with each other. For example, one transducer that encodes spelling alterations (such as the use of unaccented characters) can be composed onto a transducer encoding a lexicon to allow additional access to a unique lexical source. The practical advantage is that the finite-state rules can be compiled efficiently on a computer into a data-structure - a transducer. The transducer is a finite-state machine which consumes input while producing output. Traversing the data structure transforms the input. These transducers incorporate the context in which the transformations take place, eliminating the need for specifying programming decisions in some type of programming language, and making NLP processing relatively platform-independent.

The tokeniser

One of the very first steps in any natural language processing system is to apply a tokeniser to input text. A *tokeniser* segments an input stream into an ordered sequence of tokens, each token corresponding to an inflected word form, a number, a punctuation mark, or other kind of unit to be passed on to subsequent processing. Though most sequences of uninterrupted alphabetic characters compose a token in most languages, the use of separators inside words varies from language to language. For example, the sequence *l'amour* might split into two tokens in French, while *aujourd'hui* might be considered as a single token. On the other hand, in certain cases a sequence of words (e.g. *ein bisschen*, *a priori*, *e. g.*, *parce que*, *a fuera de*, *in order to*) may be considered as a single token for further linguistic treatment.

Our approach to tokenisation is to provide a cascade of language-dependent finite-state transducing tokenisers. These tokenisers segment text by introducing a token boundary (usually a new line) into the output stream. The cascade is composed of a basic tokeniser which segments any sequence of input characters into simple tokens (i.e. no multiword units) and one or several multiword staplers which identify multiword expressions and group them together as single units. The development and implementation of a finite-state longest match operator has made this development both practical and possible.

Morphological analysers

Now that the computer has the means to determine what a word is, its next task is to analyse words as they appear in a text.

Morphological analysis is the process which takes the surface form of a word and returns its lemma together with a list of morphological features and parts of speech. For instance, for the French surface form *lui* the morphological analyser returns the three following possibilities:

- lui+Dat+InvGen+SG+P3+PC
- luire+PaPrt+Verb
- lui+InvCase+Masc+SG+P3+PToni+Pro

The first line gives the clitic pronoun interpretation of the lemma *lui (je lui donne un livre (I give him/her a book))* together with a list of morphological tags carrying information on the fact that the surface form is invariant in gender (InvGen), singular (SG), and third person (P3), as well as information about the case

(Dat). The second line gives the verb interpretation *(la lumière a lui (the light did shine))* of the lemma *luire*, together with morphological information about the surface form: it is a past participial form of a verb. The third line gives the pronoun interpretation *(lui, qui tant de fois ... (He, who so many times))* of the lemma *lui*, together with a list of morphological information: the surface form is invariant in case (InvCase), Masculine (Masc), singular, third person, and it is a tonic pronoun.

In languages like German where the compounding process is very productive, morphological analysers not only provide morphological information, but also give suggestions on how to split words. As such they supply tokenisers for some languages. A direct advantage is that, in languages like German in which compounding is a highly productive process, they provide a means of speeding up the comprehension of compounds. For instance, no German dictionary gives *Weingärtnergenossenschaftsvorstandsvorsitzender* as an entry. Together with a list of morphological features and possible parts of speech the German morphological analyser indicates word boundaries (#):

Wein#Gärtner#Genosse\nschaft\s#Vorstand\s #Vorsitzender (wine, gardener, co-operative, committee, chief)

By indicating how to split this compound the German morphological analyser provides users with useful information about where to look in the dictionary in order to find the definition of all the pieces, put them together and eventually understand the overall meaning of the word.

Morphological analysers use finite-state technology to encode variations of words in different languages. These analysers are first created by lexicographers who describe the word classes of a language and their inflectional behaviour in declarative two-level rules, which are compiled into lexical transducers.

Part-of-Speech Taggers

Part-of-speech taggers choose the most appropriate part of speech associated with a word in a given context. Part-of-speech tagging is performed using probabilities, namely a Hidden Markov Model (HMM) of rank 1. In this model we use two probabilities: the *lexical probability* and the *transition probability*. The lexical probability is at word level. For instance, the lexical probability of the word *like* in big corpora is the number of times it appears as a verb compared to the number of times it appears as a preposition. The transition probability is at the sequence level. For instance, we compute how many times the part-of-speech sequence *pronoun preposition* appears compared to the part-of-speech sequence *pronoun verb*. The combination of these two probabilities is used to decide the most appropriate part of speech associated with a given word in a given context.

At XRCE we use the Xerox tagger. The tagsets for the different languages cover the major part-of-speech classes (nouns, adjectives, verbs, pronouns, determiners, etc.), but they differ with respect to language-specific morphological information (number, gender, inflection, etc.). The above components have been developed for seven languages: English, Dutch, French, German, Italian, Portuguese and Spanish. Currently, we are developing the same suite for Russian, Czech, Polish, Hungarian and Arabic. The same technology has also been used to produce morphological analysers for Turkish and Korean.

Incremental finite-state parsing

Finite-state parsing is an extension of finite-state technology to the level of phrases and sentences.

Our work concentrates on shallow parsing of unrestricted texts. We compute syntactic structures without fully analysing linguistic phenomena that require deep semantic or pragmatic knowledge. For instance, PP-attachment and co-ordinated or elliptical structures are not always fully analysed. The annotation scheme remains underspecified with respect to unresolved issues. On the other hand, such phenomena do not cause parse failures, even on complex sentences.

Syntactic information is added at the sentence level in an incremental way, depending on the contextual information available at a given stage. The implementation relies on a sequence of networks built with the replace operator. The current system has been implemented for French and is being expanded to new languages. The parsing process is incremental in the sense that the linguistic description attached to a given transducer in the sequence relies on the preceding sequence of transducers, covers only some occurrences of a given linguistic phenomenon and can be revised at a later stage.

The parser output can be used for further processing, such as extraction of dependency relations from unrestricted corpora. In tests on French corpora (technical manuals, newspapers), precision is around 90-97% for subjects (84-88% for objects), and recall around 86-92% for subjects (80-90% for objects).

Applications based on NLP tools

Once these basic tools are available in a language, they can be used in a wide variety of natural language engineering applications, many of which are exploitable for multilingual corpus exploration. We describe three applications below.

Multilingual comprehension aids

One of the greatest impediments to efficient understanding of foreign texts affecting readers more or less at all levels of language comprehension skills is the appearance of an unfamiliar word or phrase, and subsequent manual searching in a hardcopy bilingual dictionary. As a response to this problem, researchers have developed LocoLex, an intelligent reading aid incorporating a machine-readable bilingual dictionary and our linguistic processing suites. In addition to using the part-of-speech disambiguator in order to directly select the dictionary entry corresponding to the part of speech used, LocoLex can recognise multiword expression patterns in order to focus the user's attention on the best translation for a word in context by a regular expression encoding of multiword expressions in the bilingual dictionary. For example the idiomatic expression *take the bull by the horns* is encoded as a regular expression which matches any sequence of adverbs, any form of the verb *take* and the surface forms of the fixed part of the expression *the bull by the horns*.

An advanced demonstration version of LocoLex called TANS (Translation Aid Network Services) exists in three versions: a toolkit version that enables other programs to use its functionalities, an add-on to Word for Windows, and a version accessible through any browser on the WWW. The preliminary version, installed at the Grenoble laboratory in September 1995, includes a French to English dictionary with 40,000 entries, 11,000 idioms and 5,000 multiword expressions.

Cross-language information retrieval

As corpus access becomes more distributed and internationalised, encountering multilingual corpora during an information retrieval task will become more common. Beyond merely accepting extended character sets and performing language identification, the text retrieval systems of the future will have to provide help in searching for information across language boundaries. At Xerox Research Centre Europe, we have begun a series of experiments to explore what factors are most important in making multilingual information retrieval systems work. Preliminary results demonstrate the necessity of recognising and translating multiword units. For example, the French expression for *insider trading* is *delit d'initie*, and simple word-based translation methods will miss the correspondence between the terms.

An online implementation of some aspects of cross-language information retrieval was developed by the Callimaque project, a collaborative project led by IMAG (Institut de Mathemathiques Appliquées de Grenoble), INRIA (Institut National de Recherche en Informatique et Automatique) and XRCE Grenoble. Callimaque offers cross-language access over the Internet to a collection of 3,000 French documents showing the evolution of applied mathematics and computer science in France over the last 40 years. Scanned documents were OCRed and indexed in French using the XRCE NLP suite to lemmatise and extract indexing terminology. These innovative tools help non-French speakers to access this set of French documents. Thus, readers with little knowledge of the French language will be able to formulate a query to search the database in either French or English and capture the linguistic variations of the multiword expression they are looking for, as well as to obtain a contextual transla-

tion from French to English of certain critical pieces of text, such as the title or the abstract of a document.

Terminology extractors

Much of the terminology found in a corpus is composed of noun phrases. One extension of our NLP suite is a noun phrase extraction step which can follow part-of-speech tagging. In order to perform this step, transducers have been compiled from finite-state expressions which are basically grammar rules describing the contour and patterns of noun phrases for each language for which a lexicon and tagger are created. The patterns can include surface forms as well as part-of-speech tags. When these transducers are applied to tagged text, noun phrase boundaries are inserted.

The current noun phrase mark-up was designed basically for terminology extraction from technical manuals. It covers relatively simple noun phrase detection, i.e. some constructions such as relative clauses are not included.

Because one can easily add a new regular expression to handle more constructions, more elaborate patterns including verbs can be extracted. The same automatic means have been used to extract collocations from corpora, and in particular support verbs for nominalisations . In English, an example of proper support verb choice is *one makes a declaration* and not *one does a declaration*. *Make* is said to support the nominalisation *declaration* which carries the semantic weight of the phrase.

We used NLP suites, followed by syntactic pattern matching that was slightly more complicated than the noun phrase extractors of the previous section, to extract verbal categorisation patterns for around 100 nominalisations of communication verbs in English and French. Both noun phrase and verb phrase extractors turn out to be very useful tools for translators. Indeed, when a translator is given a new technical text to translate, the first task is to build a lexicon with the appropriate terminology. Terminology extractors enable lexicon building for various languages, either singly or on a bilingual basis.

Conclusion

Multilingual language processing necessitates a coherent range of linguistic tools, performing the same functions across languages. The XRCE-MLTT approach we presented here produces tools and techniques which are robust, as well as being applicable to large quantities of text in different languages. The technology used ensures our capability to build ever more powerful tools.

F. Segond, G. Grefenstette and A. Zaenen
Xerox Research Centre Europe
6, chemin de Maupertuis, 38240 Meylan
France
{grefen,segond}@xrce.xerox.com
http://www.xrce.xerox.com

The APOLLO Project - Achievements and Conclusions

Guy Deville and Pierre Mousel

This paper describes the achievements of the EC co-funded LE APOLLO project. The ultimate aim of APOLLO was to provide an open workbench for multilingual document creation and maintenance to the banking and finance sectors. However, the project was considered as the initial phase of a process that was designed to result in a fully-fledged version of the above-mentioned workbench. In this article, we first sketch the workbench prototype architecture and its components, then outline the results of the user survey. In conclusion, we discuss the reasons for not pursuing a follow-up of APOLLO beyond the end of the preparatory action project.

APOLLO (reference number LE-1033) is a project that was co-funded by the European Commission within the Telematics Applications Programme of the Fourth Framework Programme. APOLLO had three short-term objectives: (i) to clearly identify and specify end-user needs in the sectors of banking and finance with respect to multilingual document creation and maintenance, (ii) to develop a mock-up workbench demonstrating the possibilities of state-of-the-art multilingual document management, and (iii) to come up with a workplan study for a fully-fledged workbench.

Originally, the APOLLO project was aimed at employees executing core banking functions in situations requiring multilingual skills. The APOLLO workbench aimed to reduce the delays inherent to centralised translation workflows.

Architecture of the APOLLO Workbench

The mock-up of the APOLLO workbench relied as far as possible on existing technology. It included a number of interacting tools offering services through well-defined interfaces and consisted of (i) a text processing tool (Interscript), the main function of which was to allow users to physically layout multilingual texts; (ii) a machine translation component (CAT2), the main function of which was to translate unformatted texts from one language into another; (iii) a dictionary tool, the main function of which was to offer access to multilingual thesauri and to translate words, and (iv) an SGML-based version management tool, the function of which was to maintain successive versions of multilingual documents.

The various components were integrated in a consistent system that was available to users working on Windows PCs in a networked environment. It was user-friendly - the user interface was a WYSIWYG text processor running under Windows that implemented standard office document text formatting capabilities. It was also open - the workbench was designed in such a way that extending it with additional or alternative components would be very easy.

Implementation Issues

Text Processing

As mentioned above, the core element of the workbench was the Windows version of the text processor Interscript, a user-friendly application that runs on various platforms. Interscript offered all the functions that users currently expect from this type of product: editing, font control, search facilities, style definition, etc. We chose the Interscript text processor because we were granted access to its source code. This was mandatory as the text processor's document architecture had to be redesigned to implement multilingual documents. Indeed, in the APOLLO workbench, multilingual documents were complex objects that consisted of several subdocuments, each subdocument being a different linguistic version of the same text. The parts of a subdocument were related to their corresponding equivalents in the other subdocuments. We could not have implemented these relationships by developing an add-on to a mainstream text processor without access to its source code. Indeed, extensions (e.g. Eurolang Optimizer) to standard text processors (e.g. Microsoft Word) simply augment the application's function set without modifying its core document architecture. With such extensions, the documents basically remain monolingual documents.

Machine Translation

- CAT2 as MT Engine

CAT2 is a unification-based machine translation system developed as a sideline of the CEC-sponsored EUROTRA program. It was integrated into the text processing component with standard client/server technology. CAT2 is a rule-based system that consists of the CAT2 engine (software) and the CAT2 lexicons and grammars (lingware). For the APOLLO project, the engine was used unchanged while new lingware was developed specifically for the pilot application. The formal properties of the system can be summarised as follows. Unification is the only computational

mechanism used, and it works on the basis of tree structures and feature structures annotated to every node of the tree. Unification may be constrained by negative, disjunctive or implicative constraints over simple and complex features. In the APOLLO project, the CAT2 system was tested on a sample of bilingual (French-English) specialised texts from the sectors of banking and finance. The lexicons and grammars were developed in a corpus-based approach, as discussed below.

• Specialised Corpus-Based Lingware Development

The development of the APOLLO lingware combined a classical corpus-based approach (relying on textual data provided by the APOLLO consortium) with the reuse of general-purpose lexicographic resources.

Thus we collected and formatted 8,000 texts from various banking user bodies (specialised courseware) and a scientific documentation centre (paper abstracts). On the basis of the collected corpora, we modelled the domain sublanguage in a two-step approach. We identified the terminology in the selected courseware by using existing in-house term banks; then we linked the terminological description to general semantics in order to build the application-specific lexicons and identify particular language constructions, mainly collocations, compounds and idiomatic expressions, that had to be reflected in the grammar. Finally, the lexicons and sublanguage construction were turned into the CAT2 formalism.

Dictionary Look-up Facility

Beside the machine translation component, we also developed a dictionary look-up facility as an extension to the text processor. The look-up module's dictionaries were completely independent from those that were used by CAT2 and were implemented as a Microsoft Access database. In order to increase the module's independence with respect to a specific RDBMS, the module accessed the dictionary exclusively via ODBC. The dictionary look-up facility provided information which was quite similar to that in a paper-based dictionary. For a word in a source language, the user could get the possible translations in various target languages and choose a translation on the basis of a definition of the word. Each translation was illustrated with examples, etc.

SGML-Based Version Management

The last component of the workbench was a version manager. The APOLLO Version Manager is based on the rcs utility and is implemented as a server running on a UNIX system. For the purpose of the APOLLO workbench we designed an SGML Document Type Definition (DTD) that allows us to represent formatted multilingual documents without loss of information. The text processor Interscript was extended with conversion utilities that are able to convert multilingual documents from the Interscript format to an APOLLO DTD-compliant format and vice versa. To integrate the Version Manager, we used the same standard client/server technology as for the machine translation component.

Market Study

Aside from the development of a mock-up workbench, the APOLLO consortium also conducted a market study, the objective of which was to clearly identify and specify end-user needs in the banking and finance sectors with respect to multilingual document creation and maintenance.

The user needs study carried out in Luxembourg, Belgium and France had two facets, a qualitative and a quantitative one. For the qualitative study, we interviewed fifteen people, while for the quantitative study we evaluated answers from roughly fifty people (out of several hundred questionnaires mailed). We surveyed both the banking environment (mainly) and industry (for comparative purposes). The study revealed several interesting facts, which are outlined below.

First of all, the study highlighted important differences in translation needs, in terms of volume, between Luxembourg on the one hand, and Belgium and France on the other. Indeed, translation needs in Luxembourg seemed to be far less than those in Belgium and France. The reason could be the very particular linguistic context in Luxembourg, where almost every managerial staff speaks three major languages. This hypothesis was confirmed by the fact that none of the banks we surveyed had a specialised translation department. In Belgium, however, the volume of translations to be done quite often justified the existence of specialised departments that were sometimes of an impressive size (up to twenty people).

The market study also produced evidence of clear differences between the banking and the industrial sectors. While the banking sector required very high translation quality standards, draft quality translations were sufficient for most industrial applications. The reason for this was the difference in types of documents that were translated in both environments. In the banking field, these were mainly marketing documents, meeting reports and training handouts, whereas technical documentation is by far the most frequent document type handled in industry. Technical documents were mostly written in rather simple language that used a highly specialised but nevertheless almost entirely monosemic vocabulary. This was less the case in banking documents that used much more complex language with polysemic words because of the wide potential variety of subjects.

Finally, the last part of the market study examined the current degree of automation of the translation process in banks. Again, significant differences appeared between Luxembourg on the one hand, and Belgium and France on the other. As a matter of fact, several translation departments in Belgian banks already use a number of computer-based translation tools (from terminological databases to translation memory systems). In Luxembourg, however, none of the banks examined used any similar products. Version management was still done manually and was not computer-aided in either Luxembourg, Belgium or France.

Thus the market study clearly showed that to support the translation process automatically one has to take into account (i) the precise context within which the systems will be implemented and (ii) the specific needs of the end users. To date, there is no generic computer-based solution.

Conclusion

We could indeed identify needs in the banking sector, but these were so specific that they will most certainly never generate an important market. Our market study showed that the proposed solution could only fill small niches in the banking and finance sectors. This was confirmed by the only weak commitment which most banks were ready to make when asked to participate in a follow-up project. Although this might be due to the low awareness of language technology in banking circles, our study indicates that the main reasons for this state of affairs are more likely to be the limited capabilities of current language technology, and more specifically machine translation. The overall unsatisfying performance of machine translation technology in this field stems from the immense variety of texts to be translated in banks. Furthermore, users quite often require very high quality translations. As a consequence, the fact that machine translation needs extensive post editing, heavily reduces the benefit of using automated solutions.

We concluded that there was no strong need for a document creation and management system as conceived in the APOLLO project within the banking and financial sectors. These findings and the muted reactions of the APOLLO user group led us to drop the idea of a follow-up project.

For more information, please contact:
Guy Deville
Facultés Universitaires Notre-Dame de la Paix, Namur
Belgium
Guy.Deville@fundp.ac.be
or
Pierre Mousel
Centre de Recherche Public - Centre Universitaire
Luxembourg
Pierre.Mousel@crpcu.lu

French Government Launches "Information Society" Action Programme

The French government recently launched an action programme entitled "Preparing France's Entry into the Inform@tion Society." Billed as a document "mark[ing] the Government's commitment" to establishing France as an information society, the Action Programme comprises an outline of issues and priorities, as well as a set of proposals for government action. The Programme describes six main priorities: new IT and communications tools in education, cultural policy, modernization of public services, IT in the private sector, meeting the challenges of industrial and technical innovation, and encouraging effective self-regulation of new information networks. The government would also like the Programme to be a starting point for a wider public debate on this topic. The text of the action programme mentions ELRA in the context of the distribution of multilingual resources in close cooperation with Délégation Générale à la Langue Française (DGLF):

"Making available automatic linguistic resources is an essential condition for the development of a large number of software packages, applications and interfaces requiring language analysis. The rise of the Internet has emphasised the importance of research and indexing tools, resources of which there are still too few in French-language form.

The DGLF will lend its support to the production and distribution of multilingual resources in which French is one of the languages, in the context of the “Multilingualism and the Information Society” programme set up by the European Commission. It will back up the actions of the European Language Resources Association.

The Ministry of Culture and Communication will implement a specific initiative to clarify user rights for research scientists in certain existing bodies, such as the “Institut national de la langue française” (National French Language Institute), the CNRS, or the National library."

For more information, visit the following Web site:
http://www.premier-ministre.gouv.fr

NODALIDA ‘98

A Report from Bente Maegaard

The 11th Nordic Conference on Computational Linguistics took place in Copenhagen 28-29 January 1998. The conference was organized jointly by University of Copenhagen, Department of General and Applied Linguistics, and Center for Sprogteknologi, Copenhagen. Nodalida (NOrdiske DAtaLIngvistikDAge) is a biannual event, organized in the Nordic countries, last time in Helsinki, Finland, and next time in Trondheim, Norway. (Nordic countries: Denmark, Finland, Iceland, Norway, Sweden).

This time, Nodalida attracted 57 participants, researchers from the academic as well as the commercial world, and students.

The programme had a broad coverage of computational linguistics, spanning from very applied to very theoretical issues. The first day started with MT (IBM’s LMT system) and multilingual term bases. It went on with computational lexica, an HPSG paper on determiners and clausal adverbials, papers on parsing, and a large amount of papers on corpora and corpus linguistics. A couple of papers discussed the use of the www, e.g. for computer assisted language learning. The conference ended with a paper on the empty string in an LFG-like feature structure grammar formalism, and a paper on advanced computing in the humanities.

A panel discussion on The Nordic languages in the Information Society - a responsibility for computational linguistics and computational linguists? showed that the interest from public funding agencies in supporting computational linguistics has been and is different in the various Nordic countries. The Nordic Council discussed the language issue last May, and it was suggested that a coordinated approach be made to the Nordic Council concerning the protection and reinforcement of the Nordic languages which are ‘less used’ on a world basis.

For more information, please contact:

Bente Maegaard
Center for Sprogteknologi, Njalsgade 80, 2300 Copenhagen
Denmark
Tel: +45 35 32 9074
Fax: +45 35 32 9089
E-mail: bente@cst.ku.dk

Q&A - ELRA members

Starting in this issue, we are presenting a new feature: Q&A - ELRA members. This will be ongoing, highlighting our members in short profiles such as the ones found below. If members wish to be featured, please contact the ELDA office on +33-1-43 13 33 33 or elra-elda@calva.net.

CLIF (Research Community for Computational Linguistics in Flanders), Belgium

- What are the main activities of your organisation?

Members of CLIF are all university research groups.

- Which is the main interest out of speech/text/terminology for your organisation?

Members of CLIF are mainly involved in text and speech research.

- Why are your organisation an ELRA member?

In the first instance, CLIF was interested in linguistic resources. An urgent need was felt to be able to acquire (text) corpora, databases, and the like.

- What do you expect from ELRA in the future?

We would like to see ELRA develop an active policy in acquiring annotated text and speech corpora and user friendly linguistic databases, so that it can show its genuine complementarity to LDC.

Ericsson Mobile Communications AB, Sweden

- What are the main activities of your organisation?

Voice Communication, Telephony.

- Which is the main interest out of speech/text/terminology for your organisation?

Speech. We use Speech resources for testing, training and for voice algorithms.

- Why are your organisation an ELRA member? What do you expect from ELRA in the future?

The need for speech and related databases in industrial and research laboratories will increase with the globalization trend in the world. I see that ELRA can play an important role in providing such databases and by becoming a link to other international speech databases.

5TH INTERNATIONAL CONFERENCE ON SPOKEN LANGUAGE PROCESSING - ICSLP '98

Sydney Convention & Exhibition Centre November 30th - December 4th 1998

ICSLP '98 offers Keynote presentations and other plenary events which bring both experience and vision of multi-disciplinary attacks on grand challenges in spoken language processing in both humans and machines. A student day at which full-time student registrants may present their ideas under the guidance of senior mentors is also planned. However, it is the quality of the delegate presentations which will be the major factor in making ICSLP '98 a truly landmark event. We invite you to attend.

CONFERENCE SECRETARIAT:

Tour Hosts Conference & Exhibition Organisers
GPO Box 128
SYDNEY NSW 2000
AUSTRALIA
Tel: +61 2 9262 2277 - Fax: +61 2 9262 3135
Email: icslp98@tourhosts.com.au

IMPORTANT DATES:

- Friday 1st May, 1998
  Paper summaries due for review
- Friday 26th June, 1998
  Acceptance notification
- Friday 21st August, 1998
  Deadline for full-paper submission

Information on the areas of submission and how to submit can be found at:

WWW = http://cslab.anu.edu.au/icslp98 - E-mail Submission: icslp98@one.net.au
Postal: ICSLP '98 Secretariat, GPO Box 128, Sydney, NSW 2001, Australia
Technical queries: Robert Dale - email: rdale@mpce.mq.edu.au • General Information: Email: icslp98@tourhosts.com.au

New resources

ELRA-S0049 The SPK database

SPK is an Italian speech database of isolated and connected digits. It was designed and collected at the Istituto per la Ricerca Scientifica e Tecnologica (ITC/IRST), Trento, Italy. SPK was conceived for speaker recognition and verification purposes.

With this CD-ROM, speech material corresponding to isolated digits acquired from 100 speakers (30 females and 70 males, from 23 to 50 years old) is released. Most of the speakers are from the North-East of Italy. Speech material was collected from each speaker during five recording sessions scheduled on different days. During a recording session four repetitions of the ten Italian digits were acquired from a speaker. A total of 20,000 speech waveform files form the corpus.

Recordings were performed in a quiet room. Speech was acquired at 48 kHz, with 16 bit accuracy, by means of a Digital Audio Tape-Recoder Sony TCD-D10PRO and a super-cardioid microphone Sennheiser MKH 416-T. Then, digital recordings were downsampled to 16 kHz. Speech waveform files in the corpus were stored in the NIST-SPHERE format by using the SPHERE library, version 2.6a.

Price for ELRA members:
for research use: 400 ECU
for commercial use: 800 ECU

Price for non members:
for research use: 800 ECU
for commercial use: 1,600 ECU

ELRA-L0030 Bilingual Collocational Dictionary (Horst Bogatz)

The bilingual English-German collocational dictionary consists of around 40,000 English headwords, including concepts expressed by more than one word (e.g. "environmental awareness" or "lame duck") and hyphenated compounds. It contains verbs, adjectives, synonyms and phrases that collocate with the headword, as well as the German equivalents for the headwords and their English synonyms.

The corpus on which the dictionary is based consists of a representative group of written (British) English texts books, magazines, and quality Press which runs to about two million words. All entries are based on contemporary evidence, and are typical of words that appear at least once in a two-million word corpus. The examples and phrases are a major feature of this dictionary.

A global search will provide all collocations that can possibly be associated with the search word. A search engine, the Advanced Reader's Collocation Searcher (ARCS), is supplied with the data and provides all possible German equivalents of the headwords. All entries are sorted according to part-of-speech categories. The latter feature makes it possible for searches to yield different useful combinations of words, e.g. noun + verb + adjective + examples extracted from the corpus + synonyms. A global search will also locate all words semantically connected with the search word in both English and German.

File format: 8-bit ASCII
Medium: CD-ROM
Price for ELRA members: 210 ECU
Price for non members: 300 ECU

ELRA in the News

ELRA was the subject of an article by Bernard Montelh in Le Monde - Radio, Télévision, Multimédia - of Sunday 1 February 1998. An abstract of the text has been translated and is reproduced below (courtesy of Le Monde):

"...Many resources are lying around unused in research laboratories when they could be used. In European projects, such resources are lost when the project ends", says Khalid Choukri, CEO of the European Language Resources Association (ELRA). Set up at the end of 1995 on the initiative of the European Commission, the association is incorporated in Luxembourg but based in Paris. Its object is precisely to identify those resources which might interest public sector organisations and private enterprises working in the field of language, to negotiate the rights with their producers and to ensure their distribution. "The rise of the Internet and, more generally, international changes are leading to a growing demand for sophisticated search tools, online dictionaries, machine translation systems, spelling and grammar checkers and gisting tools", explains Khalid Choukri. "To work, all of these tools have to be based on major linguistic corpora which may either be general, or specialised e.g. for translating technical documentation."

In addition to written resources and terminology, ELRA is also looking for speech data, which is of interest in relation to speech recognition research (dictation, voice navigation, human-to-machine telephone dialogs, etc.). Researchers in this area need the most representative samples of real-life speakers possible. "The best example is Lemsi (sic), a CNRS laboratory, which created the first French oral database. We were certain that it would be of interest to quite a lot of people, so we convinced the researchers and then the CNRS's legal service to license it in return for royalty fees."

One of the main barriers to the dissemination of data outside its original context are legal and financial considerations. The texts must be free of all rights and encumbrances, which is not the case with contemporary works. This means that only some of the 3,500 texts dating from the sixteenth century down to the present day which are contained in the corpus used to build the dictionary known as the Trésor de la langue française, a national dictionary established by the INaLF (Institut National à la Langue Française), could be ported to the Internet (restricted access is available via the site hosted by the INaLF… and unrestricted access via the University of Chicago's Website!). The electronic version of the dictionary (the TLF itself), which uses a Web navigator as an interface, currently only comprises five volumes. Six others will be ready before summer and the last six (chronologically speaking the oldest six, which are the most difficult to handle due to the differences in the language) will be completed towards the year 2000. However, Jacques Dendien, head of IT at INaLF, hopes to be able to put up the parts already available on the Internet relatively quickly once evaluation is complete. This is the background to the tough negotiations currently going on with Gallimard, the publisher of the printed version.

"In Germany, things are clear-cut: data created by public sector research bodies using public money belong to everyone, and companies want to make the most of this. In France, the position is that the State only facilitates creation, and the research labs remain in possession of their work and can do as they please with it", explains Khalid Choukri. "However, we have a vested interest in their dissemination, especially in order to be able to meet the needs of corpus linguistics, which are not competing with such base applications. When all is said and done, what is at stake is the position of the French language on the Internet."