The ELRA Newsletter

December 1996

Vol.1 n.4

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.

Editor in Chief:
Khalid Choukri
Editor:
Deborah Fry
Layout:
Martine Garnier-Rizet

Contributors:
Nicoletta Calzolari
Annie Coret
John Hutchins
John McNaught
Ole Norling-Christensen
Thomas Schneider
Kenji Sugiyama
Antonio Zampolli

ISSN: 1026-8200

ELRA/ELDA
CEO: Khalid Choukri
Assistant: Rébecca Jaffrain

87, avenue d'Italie
75013 Paris
Phone: (33) 1 45 86 53 00
Fax: (33) 1 45 86 44 88
E-mail:
elra@calva.net
WWW:
http://www.icp.grenet.fr/ELRA/home.html

Dear ELRA Members,

In this second of three issues of the ELRA Newsletter devoted to the different Colleges, we have turned our attention to the field of Written Resources. In addition to providing in-depth articles on particular areas such as corpora and machine translation, we can report on a number of practical projects and implementations of research results, including AVENTINUS, LE-PAROLE, and EAGLES.

Moving beyond this theme, the issue also provides information on a number of major "internal" events, such as the second Annual General Assembly, held in Paris in December 1996, and the arrival of ELDA's second technical assistant, Valérie Mapelli. It also features updated lists of members and resources. The latter document in particular shows the progress which ELRA has made towards fulfilling its objective of acquiring and making available reusable resources to the language engineering industry. This was confirmed in the positive review by the European Commission in early December. We hope that these measures have laid a sound foundation for ELRA's third calendar year of activities, and for its long-term success.

In 1997, ELRA will be shifting its focus slightly to strengthen its marketing and distribution activities. A meeting of the Expert Panel on Marketing and Distribution, to be held in Mainz in January, will examine new distribution options and methods of enhancing ELRA's visibility in its different target markets, plus more sophisticated pricing mechanisms reflecting the extremely heterogeneous language resources now being offered to ELRA for distribution. Pricing policy has been already discussed in several forums and a document is now available on the web (on the members only section) which intends to stimulate an internal debate as a prelude to finalising the official policy of the association. In conjunction with this debate, the LR description forms have been revised to take the feedback of our members and experts from several panels into account (these forms are also on the Web).Also on the Panel's agenda are ways in which ELRA can increase its value to members, e.g. through further improvements to the Catalogue and additional services on ELRA's Web site, which is updated regularly.

One concrete example of our new policy will be the survey of members planned for the spring and designed to elicit people's needs, wishes and dislikes as a basis for future policy. Another is the progress we are making on validation, where several sub-contracts have now been awarded. A first draft of a validation manual for lexicons will be ready in the very near future and will be made available to our members for comments.

Other measures will be implemented over the next few months and will be reported on in more detail in this Newsletter, among other places. As always, we would welcome any comments or suggestions for improvements that you may have in any areas. Similarly we would like to repeat our invitation for members to offer us their resources for distribution, or to make us aware of resources held by non-members which we might obtain.

With best wishes,

Antonio Zampolli, President Khalid Choukri, CEO

Valérie Mapelli - ELDA Technical Assistant

Born in 1972 in Annemasse, France, Valérie Mapelli studied international trade and languages at the University of Savoy, before obtaining a master's degree from the European Business Management School at the University College of Swansea (Wales) and a post-graduate diploma in applied computational linguistics from the University of Metz, France. After working on the CRIN DIALOGUE team's SILFIDE interactive server for French language resources, and (during her studies) on the Italian module of the GALEN (Generalised Architecture for Languages, Encyclopedias and Nomenclatures in medicine) project at the University Hospital in Geneva, she has now joined ELRA, where she manages ELRA contributions to its French-funded projects. In addition, she is responsible for maintaining the ELRA catalogue and for the content of the ELRA Web site.

ELRA Board Profiles

George Carayannis

ELRA founder member George Carayannis, Director of the Greek Institute for Language and Speech Processing (ILSP) and President of the Greek Pedagogical Institute, has been a professor at the Computer Science Division of the National Technical University of Athens (NTUA - Department of Electrical Engineering) since 1984. Before that, he headed the Greek research unit for the EUROTRA machine translation project.

After graduating in electrical engineering from NTUA in 1969, he obtained degrees in nuclear reactor physics and computer science from the University of Paris in 1970 and 1971 respectively, plus doctorates in computer science (Paris 1973) and physics (Orsay 1978).

Following work with the Ecole Nationale Supérieure des Télécommunications in Paris (1971 - 1974) and Brussels Free University (1974 - 1978), he was a project manager for speech analysis technology at the National Defence Research Centre in Athens. Joining the NTUA in 1980 as a research fellow, he was Secretary of the European Joint Committee for Scientific Co-operation of the Council of Europe from 1981 to 1983. He also sat on the ESPRIT and IT&C Management Committees from 1990 to 1992, and is a member of the Language Engineering Working Party of the Telematics Management Committee.

The co-ordinator of many national and European language and speech processing projects, Professor Carayannis has evaluated numerous EU research projects and programmes and published more than 120 papers on subjects such as digital signal processing (DSP), computational complexity and fast algorithms in DSP, speech analysis and synthesis, biological signal processing, image and speech recognition, multimedia information systems, linguistic processing and educational technologies. He has advised the Greek President's Office on the introduction of new technologies to the public sector (1994-1995), and is a founder member of the Greek Technology Assessment Association (1995), as well as a member of the Greek Chamber of Technology, of the IEEE and of EURASIP. He is also a long-serving member of the editorial board of the European journal "Signal Processing".

Giuseppe Castagneri

Born in 1954, Giuseppe Castagneri studied experimental psychology at the University of Padua and is currently Senior Researcher in the Voice Application and Services Group of CSELT's End-User Service Department. Following initial work in the field of pyschoacoustics as applied to telephony, he now specialises in the human factors involved in speech-based human-computer interaction.

Giuseppe Castagneri participated in the ESPRIT SAM and SAM_A projects on speech recognition assessment, and has also been active in the field of database collection. He organised and chaired the Workshop on "International Co-operation and Standardisation of Speech Databases and Speech I/O Assessment Methods" in Chiavari (Italy), and was involved in LRE project 62057 (EUROCOCOSDA) and in the SPEECHDAT projects. A member of the Italian Commission of Speech Databases (AIDA), he also co-ordinated the collection and printing of several speech databases.

His research interests include the human factor aspects of human-computer interaction, with particular emphasis on voice interaction. In this field he co-ordinated the RAILTEL project and the Italian Consortium in the ARISE project.

Giuseppe Castagneri, who represents CSELT in ELRA, was elected to the Board at the first General Assembly, and has since been nominated a member of its Ethics Committee. He believes that ELRA's role in the speech area is to provide European scientific and industrial laboratories with the material and tools they need in order to compete internationally on the quality of their studies and products.

Harald Höge

Born in Hirschberg/Saala (East Germany) in 1945, Harald Höge studied physics at J.W. Goethe University in Frankfurt am Main, before starting work in 1970 at Siemens' Central Communication Laboratory. Since 1978, he has headed a speech processing team (currently comprising 19 people) working on speech recognition, speech coding and speech synthesis. Applications include speech processing for telephone switches (EWSD, HICOM), for terminals (cellular phones and cars) and for PCs (the "hearing typewriter").

Harald Höge has organised and/or participated in numerous national and international projects, the most important of which are SPICOS (a joint project with Philips which developed the first speech dialogue system handling continuous speech); VERBMOBIL (a large project for speech-to-speech translation funded by the German government); C-STAR (an international speech translation project which gave its first intercontinental demonstration in 1994 for German, English and Japanese), and SpeechDat (an EU project designed to collect speech databases for telephone applications).

Awarded a doctorate in 1994 for work on adaptive echo cancellation in long distance communication, Harald Höge was also vice-president of ICASSP96, and is a member of GI, ELRA, BAS and ELSNET. Other activities include guiding work on the development of speech processing algorithms for commercial applications, co-operation with universities and the supervision of doctoral theses in an industrial environment, setting up projects promoting multilingual approaches, and theoretical work on the entropy of the information content of speech.

Habeas corpus

Ole Norling-Christensen

The original meaning of the Latin word corpus (plural corpora or, in English, also corpuses) is a human or animal "body", be it dead ("corpse") or alive. However, Cicero and other Roman authors already used corpus, or its Greek equivalent σῶμα, as the technical term for "a whole composed of parts united" (Lewis and Short[1]), and especially for a "complete collection of writings about a specific topic". Corpus Juris, for instance, is the comprehensive collection of Roman Law which was compiled in the 6th Century[2]. It is only in the second half of our century that a new sense of corpus - the one which is relevant in the context of ELRA - has emerged, namely "the body of written or spoken material upon which a linguistic analysis is based" (Oxford English Dictionary, 1956), "ensemble limité des éléments (énoncés) sur lesquelles se base l'étude d'un phénomène linguistique" (Robert, 1961), or "Sammlung einer begrenzten Anzahl von Texten, Äußerungen o. ä. als Grundlage für sprachwissenschaftliche Untersuchungen" (Duden, 1974).

It was during this period that computers became available and, following the pioneering work of Busa[3], computational methods and techniques were gradually adopted by linguists, especially for major lexicographic projects. Since then, corpora have become a primary source for lexicon building and grammar engineering, as well as playing an important role in the training of statistical language models, e.g. for part-of-speech tagging, and in the evaluation of tools for natural language processing. Furthermore, a number of dictionaries for human users are currently being constructed on the basis of corpora.

What is a corpus?

In one of its final reports, the MLAP PAROLE project[4] gives the following definitions of terms relating to corpus linguistics[5]: An (electronic text) archive is "a repository of readable electronic texts not linked in any coordinated way". All kinds of text, but normally entire texts rather than samples, may thus be present in an archive. They may be stored on all kinds of digital media and in all kinds of formats, such as the proprietary formats of different word-processors and computer typesetting devices, but also TEI-compliant SGML. An example of such repository is the Oxford Text Archive. An (electronic text) library (or ETL, French textothèque) is "a collection of electronic texts in standardised format with certain conventions relating to content, etc. but without rigorous selectional constraints". Typically, such a library would be derived from the archive by conversion of a selection of items from it to a standard format such as SGML. A corpus is "a subset of an ETL, built according to explicit design criteria for a specific purpose"; while, finally, a subcorpus is "a subset of a corpus, either a static component of a complex corpus or a dynamic selection from a corpus, made during on-line analysis".

The hierarchy of archive / library / corpus / subcorpus, which is partly based on work by Bernard Quémada[6], was chosen because it, like the EAGLES[7] corpus typology, clearly underlines the fact that a corpus, in the new sense of the word, should be a consciously made selection of linguistic material compiled with a specific purpose in mind. In contrast, the older sense only presupposes that the collection is complete, and no specific purpose is implied. The collected works of Shakespeare may thus be regarded a corpus in the old sense of the word, although in the new sense, it only qualified insofar as it is used (or intended) for studying his authorship.

The following definitions of different types of corpora were taken from EAGLES and NERC[8]: A specialised corpus is a task- or sublanguage-oriented collection of one or more selected sublanguages. A general-purpose corpus is a collection of a broad variety of written and transcribed spoken material reflecting language variety. A monitor corpus is a large and dynamic text corpus, which is continuously updated by discarding parts of it (typically the older parts) and replacing them by other (newer) parts.

Finally, there is the reference corpus, which gave the NERC study its name, but which is not explicitly defined in the NERC report. According to the EAGLES corpus typology, it is "a corpus designed to provide comprehensive information about a language" - a definition which seems to cover the same concept as the NERC/PAROLE "general-purpose corpus". PAROLE states more exactly that a reference corpus should be "a general-purpose corpus that has acquired a certain definitive status with respect to a particular language at a particular time in its history".

The NERC recommendations talk about poly- or multi-functional general language corpora, which should be constructed for at least all the official languages of the Union, and which should preferably "be large and open-ended .. contain written and transcribed spoken language .. cover a broad variety of language types, "topic" being taken into account as a selection criterion .. contain full texts rather than samples .. be extensively documented .. ". They further distinguish between a core component "defined as the minimal multifunctional corpus sketched above", and a peripheral component which "contains all other textual materials stored at the national node in the European network". In terms of the definitions given above, the core component is a reference corpus, whereas the peripheral component, which may contain material obtained according to availability, or acquired for particular needs or projects, is the remainder of the node's text library or even archive. One reason for the variation in terminology is probably that the term corpus is ambiguous even in the professional linguistic community; in fact, the older sense of "complete collection" is still far from being obsolete.

Generic corpora

The rationale behind initiatives like NERC and PAROLE is the assumption of a widespread need for standardised multi-purpose corpora and lexicons - so-called generic linguistic resources. However, if we - like most experts in this field - adhere to a corpus definition which focuses on specific purpose, we might conclude that it is impossible to make such a thing as a "generic corpus". Fortunately, this is not quite true. First of all, in order to make a specific corpus, one will need an ample amount of different kinds of machine-readable texts; and the job is made considerably easier if these texts are available as an electronic text library using a well-documented standardised computational representation such as the one recommended by EAGLES. Secondly, at least one "specific purpose" is of general interest, namely a carefully selected reference corpus which can be regarded as a reliable norm for the contemporary general language. Like the geoid, which forms the basis for geodesists' measurements of the heights of mountains and depths of oceans, such corpora will define the core vocabulary and grammar of the language in question; and the features which are specific to sublanguages and/or individual texts can be measured against it by linguists and language engineers.

Documentation and annotation

According to NERC, extensive documentation per corpus text, by a uniform system of distinctive features, should be provided for all information that could have relevance for specific selections. However, the list of possibly relevant descriptive features is endless, and it will only be feasible to apply a smaller

set of the linguistically most relevant features, such as medium, genre, subject field and year of (first) publication. In addition, the relevant bibliographic information must be rendered in order for the users to identify the sources and apply their own criteria for text classification if needed. In the TEI/EAGLES/PAROLE model for text representation, this documentation and classification is part of the so-called header, which is separated by special mark-up from the text proper.

The mark-up inside the text is of two kinds. Some inherent properties of the text, such as its segmentation into chapters, sections, paragraphs, etc. will be expicitly marked, whereas others, like the line and page breaks of the original printed text are normally regarded as unimportant. On the other hand, the corpus builder may enrich the text with grammatical annotations such as parts of speech and phrase structures. This latter kind of annotation, known as tagging, may be done automatically, but only if the tagger has been trained, or the effects of its rules have been checked against a manually-corrected reference. At least a part of the reference corpus must therefore be tagged and meticulously checked by linguists.

"You must have a corpus"

The title of this article, Habeas corpus, is the Latin opening phrase of a British law passed in 1679, which codified an important civil right and received the Royal Assent. At the time, the meaning was "You (i.e. the authorities) must bring the body (of a person into court in order to investigate your right to keep him imprisoned)". However, it may also be translated as "you must have a corpus", and what the European Commission is in fact doing by supporting PAROLE and related projects is actively working for another human right - the right of all citizens of the Union to have equal access to language resources in their own languages. Although the corpora and lexica may be used mainly by linguists and language engineers, their work will contribute to the preservation of the linguistic diversity of the Union and will thus have a direct impact on everyone's living conditions.

1 This dictionary citation from A Latin Dictionary (Oxford 1958) by Ch. T. Lewis and Ch. Short, and the following ones from The Oxford English Dictionary (OED, Second Edition on Compact Disc, Oxford 1992), Le Robert électronique (Paris, 19989), and Das große Wörterbuch der deutschen Sprache (Duden, Mannheim, 1978), were taken from Henrik Holmboe's paper Genbrug af korpora [Reuse of corpora], LexicoNordica 3, Oslo 1996.

2 Yet another meaning of corpus stems from the type used in a 16th century printed edition of Corpus Juris Civilis (the civil law): to German and Scandinavian typographers Korpus means a font size of 10 points (between Bourgeois and Cicero).

3 The first experiments on the use of computers in text analysis are ascribed to Roberto Busa S.J., who in 1951 published "Sancti Thomas Aquinatis hymnorum ritualium: Varia specimina concordantiarum" in Primo saggio di indici di parole automaticamente composti e stampati da macchine IBM a schede perforate [a first example of word indexes automatically compiled and printed by IBM punched card machines] (Bocca, Milan 1951).

4 MLAP-PAROLE was a preparatory project conducted in 1994-1995 aimed at transforming the NERC recommendations into a concrete working plan. Its continuation, LE-PAROLE (1996-1998) is producing corpora and lexica for 14 languages.

5 Corpus linguistics is the study, regardless of its theoretical basis, of language through the analysis of large quantities of naturally occurring data.

6 The hierarchy, and the four definitions, were taken from Atkins, Sue / Clear, Jeremy / Ostler, Nicholas: "Corpus Design Criteria" in Literary and Linguistic Computing 7,1:1-16 (Oxford, 1992).

7 EAGLES, the European Advisory Group for Language Engineering Standards, has published draft standards for corpus typology (EAG-CSG/IR-T1.1., Version dated October 1994 by John Sinclair), text typology, text representation and morphosyntactic annotation for corpora.

8 NERC, Network of European Reference Corpora, was a feasibility study carried out 1991-1995; its results and recommendations, which to a great extent formed the basis for PAROLE, are published in Calzolari, Nicoletta / Baker, Mona / Kruyt Johanna G.: Towards a Network of European Reference Corpora (Linguistica Computazionale, Vol. XI. Giardini, Pisa 1995).

More information may be obtained from:
Ole Norling-Christensen
Danish Parole co-ordinator
The Society for Danish Language and Literature, Copenhagen
Email: olenc@coco.ihi.du.dk

LE-PAROLE: Its history and scope

Antonio Zampolli and Nicoletta Calzolari

The main phase of the PAROLE project, launched under the European Commission's second LE Programme Call, is designed to produce an initial set of publicly available harmonised corpora and lexica for the major European Union languages. This article gives a brief overview of the origins of the project, the work to be performed and the actors involved.

The early days

"The tendency predominant in the 70s and in the first half of the 80s to test linguistic hypotheses with small amounts of (allegedly) critical data, rather than to study extensively the variety of linguistic phenomena occurring in communicative contexts" (Godfrey, Zampolli, 1997) has certainly contributed to the lack of interest in the same period on the part of the NLP sector in the creation and analysis of large corpora, and in the construction of extended lexica.

The 1986 Grosseto (Tuscany) workshop entitled "On automating the lexicon" (Walker et al., 1994) is usually recognised as reversing this trend and starting the process which has gradually led major NLP players to pay more and more attention to reusable language resources.

This process, which was supported by a number of initiatives following on directly from the Grosseto workshop, achieved a crucial breakthrough with the recognition in the so-called Danzin Report (1992) of the infrastructural role of LR. In 1991, the European Commission's DG XIII had asked a panel of experts chaired by A. Danzin to produce a strategic document delineating the general framework, benefits, main objectives, priorities and organisational and financial conditions needed for the development of the language industry in Europe. The final Report adopted Antonio Zampolli's suggestions that such development could only be based on the facilities and conditions provided by a dedicated European infrastructure, that the establishment of this infrastructure should be the responsibility of European and national authorities, and that adequate, reusable language resources are a central component. The Danzin Report has since had a strong influence on the formation of the Commission's current strategy, with the result that language resources are now regularly included in EC initiatives in the language processing field.

Like a number of other current actions, including EAGLES and ELRA itself, the PAROLE project ultimately derives from initiatives proposed at the Grosseto workshop. The Council of Europe, one of the co-sponsors of the workshop, set up a group of experts from European institutes

with a well established tradition in the field of lexical and corpora studies. This body was to explore the feasibility of harmonising their activities, and establish a Network of European Reference Corpora (NERC)[1]. The group, which was gradually extended to include members of all European Union languages, in turn constituted the PAROLE Consortium, which is now performing the LE-PAROLE project.

The PAROLE project

The current Consortium members are Pisa Ricerche (the co-ordinator); GSI-Erli; Institute for Language and Speech Processing (ILSP); Institut d'Estudis Catalans (IEC); University of Birmingham; Institute for Language, Speech and Hearing, Univ. of Sheffield (ILASH); Det Danske Sprog- og Litteraturselskab (DSL); Center for Sprogteknologi (CST); Institiúid Teangeolaíochta Éireann (ITÉ); Dept. of Swedish, Språkdata, Göteborgs Universitet; Department of General Linguistics, University of Helsinki; Instituut voor Nederlandse Lexicologie (INL); Université de Liège BELTEXT; Centro de Linguística da Universidade de Lisboa (CLUL); Instituto de Engenharia de Sistemas e Computadores (INESC); Fundacion Bosch Gimpera Universitat de Barcelona; Institut für Deutsche Sprache (IDS), and the Institut National à la Langue Française, CNRS (INaLF).
At the suggestion of DG XIII, the PAROLE partners are setting up co-operative networks for written language resources in their respective countries, and acting as their co-ordinating nodes. The PAROLE Association has been founded to ensure the continuity and stability of these links. It will work in close co-operation with ELRA, which will provide services for the project's validation and distribution phases.

The central goal of LE-PAROLE is to produce an initial set of harmonised corpora and lexica. A corpus of at least 20 million words and a lexicon of 20,000 lemmas will be produced for Catalan, Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portuguese, Spanish (lexicon only) and Swedish, while smaller corpora of 20, 15 and 3 million words respectively will be produced for Belgian French, Irish and Norwegian. The preparatory phase of the project, the main goal of which was to produce detailed specifications for the PAROLE corpora and lexica, was performed within the framework of the MLAP Programme (1993 - 1994).

Some of the main characteristics of the corpora and lexica can be summarised as follows:

Corpora

All the information explicitly represented in the source texts will be encoded following essentially the CES (Corpus Encoding Standard) designed by EAGLES, and on the basis of the TEI guidelines. 250,000 running words will be tagged at the morphosyntactic level following the EAGLES guidelines, and instantiated by the individual PAROLE partners for their own language.

Each corpus will be accessible for consultation, possibly via the Internet. A subset of three million words from each corpus (including the tagged words) will also be "distributable" - i.e. a physical copy of it can be given to users, and co-operation with ELRA will be sought to this end. Restrictions on the type of usage will depend on the limitations imposed by the holders of the copyright of the source texts when they authorised the inclusion of their texts in the corpus.

Each partner will construct, mark up and tag the corpus using a software package of his/her choice. The compatibility and interchangeability of the various corpora is ensured by the adoption of common criteria for composition, encoding and linguistic annotation.

Lexica

The PAROLE Lexicon model is based on the results of the LRE EAGLES and EUREKA GENELEX projects. As a result, all the lexical resources being developed are declarative, theory- and application-independent and multifunctional, as well as being able to evolve easily, e.g. to incorporate other levels of information or to become multilingual. This approach, which meets the requirements of genericity, explicitness and variability of granularity, guarantees large-scale reusability. The model - which offers a high level of precision with respect to the description - is in fact designed to ensure that application-dependent models of data and applicative dictionaries can be derived from this information repository, by mapping the application model from the generic one. 20,000 entries per language are described at the morphological and syntactic levels, and in a few cases at the semantic level as well.

The availability of rather large, uniformly structured lexical resources in all the languages mentioned above will offer users the benefits of a standardised base.

The exchange format for the lexicons - like that for the corpora - is SGML, with all the lexicons sharing the same DTD for the morphological and syntactic layers. Moreover, the use of a common set of lexicon management tools is a guarantee that all lexicons will fully conform to the model. The use of these tools is a precondition for the industrial-strength quality of the large volumes of data (in so many languages) that PAROLE is to deliver.

All the lexicons will be publicly available at conditions to be determined within the project.

The bibliography is available on the ELRA Website.

1 The NERC group produced the NERC report (Calzolari et al., 1996, and Zampolli, 1996) and organised the 1992 international Pisa meeting on corpora attended by European and North American experts, at which the ECI (European Corpus Initiative) was launched.

Computer-based translation systems and tools

John Hutchins

Machine translation (MT) is still better known for its failures than for its successes, and labours under misconceptions and prejudices deriving from the ALPAC report more than thirty years ago. The idea of developing fully automatic general-purpose systems capable of near-human translation quality has long been abandoned. The aim of MT research and related activities is to produce aids and tools for professional and non-professional translators which exploit the potential of computers to support human skills and intelligence. This research is now taking place in the context of rapid growth in the use of MT systems and translation tools, and is thus inevitably more oriented towards specific needs than some of the more idealistic research of previous decades. The following brief survey emphasises European developments.

The recent growth of MT

The traditional MT user has been the large multinational company which requires technical documentation and operating manuals in a range of languages. The system runs on a mainframe and produces 'raw' output of variable quality for revision (post-editing) by translators. A successful alternative has been the pre-editing of input texts (typically with a controlled

language) to minimise the expensive editing processes. Both these types of MT use are continuing to expand rapidly. There are now millions of pages of translation produced every year (See the reports in MT News International No.6, September 1993, and No.12, October 1995).

Although MT software for personal computers began to appear in the early 1980s (with the Weidner MicroCAT system becoming particularly successful), it has been during the current decade that sales of these systems have shown a dramatic rise. There are now estimated to be some 1,000 different MT packages on sale (when each language pair is counted separately.) The products of one vendor (Globalink) are present in at least 6,000 stores in North America alone, and in Japan one system (Korya Eiwa from Catena, for English-Japanese translation) is said to have sold over 100,000 copies in its first year on the market. Nearly all the Japanese computer and software companies (Fujitsu, Toshiba, NTT, Brother, Catena, Matsushita, Mitsubishi, Sharp, Sanyo, Hitachi, NEC, Panasonic, Kodensha, Nova, Oki, etc.) seem to have a product, mainly for Japanese and English in both directions. Outside Japan, products come mostly from independent companies set up to develop and market a range of translation software products (e.g. AppTek, CITAC, EJ Bilingual, LEC, Neocor, PC-Translator, and Globalink).

Though it is difficult to establish how much of the software purchased is regularly used (some cynics claim that only a very small proportion is tried out more than once), there is no doubting the growing volume of 'occasional' translation, i.e. by people from all backgrounds wanting gists of foreign text in their own language, or wanting to communicate in writing with others in other languages, however poor the quality. It is this latent market for low-quality translation, untapped until very recently, which is now being discovered.

Vendors of older mainframe systems (Systran, Fujitsu, Metal, Logos) are being obliged to compete by downsizing their systems: many have done so with success, managing to retain most features of their mainframe products in the PC-based versions. Systran Pro, for example, is a Windows-based version of the successful system developed since the 1960s for clients world-wide in a large range of languages; its large dictionary databases offer clear advantages over most newer cheaper products. Systran Pro is available for the language pairs (both directions) English-French, English-German, English-Spanish; and for English to Italian and Japanese to English. The publishing company Langenscheidt is now marketing a PC version of Metal for German-English translation, and shortly for other languages. The competition is clearly intensifying.

At the same time, many MT vendors are providing network-based translation services for on-demand translation, with human revision as optional extras. In some cases these are client-server arrangements for regular users; in other cases, the service is provided on a trial basis, enabling companies to discover whether MT is worthwhile for their particular circumstances and in what form. Such services are provided, for example, by Systran, Logos, Globalink, Fujitsu, NEC and MTSU (Singapore).

A further sign of the influence of Internet is the growing number of MT software products for translating Web pages. Japanese companies led the way: nearly all the companies mentioned above have a product on this lucrative market, and they were followed quickly elsewhere (e.g. by Systran, Globalink and Intergraph).

The most dramatic change of all has probably been the use of MT for electronic mail. Two years ago CompuServe introduced a trial service based on Intergraph's Transcend system for users of the MacCIM Support Forum. Six months later, the World Community Forum began to use MT for translating conversational e-mail. Usage has rocketed. Most recently, CompuServe introduced its own translation service for longer documents either as unedited "raw" MT or with optional human editing. It will soon offer MT as a standard for all its e-mail.

The use is not simple curiosity, although that is how it often begins. CompuServe records a high percentage of repeat large-volume users for its service: about 85% for unedited MT - a much higher percentage than might have been expected. It seems that most output is used for assimilation of information, where poorer quality is acceptable. The crucial point is that customers are prepared to pay for the product - and that CompuServe is inundated with complaints if the MT service goes down! It should be remembered that France was in fact the first location for a networked MT service - the networking of Systran on the Minitel system. This, too, proved popular for many different and unexpected purposes, but globally its impact has been less than CompuServe's service.

With cheaper PC software and wider access to the Internet, there has undoubtedly been an unprecedented growth in the use of MT, and primarily among non-professional translators. It should be remembered that in multinational companies the users of MT (the post-editors) are not always professionally qualified translators, although many have been specially trained to deal with MT output.

In Europe the growth has been slower than elsewhere, and PC-based systems are not yet being purchased on the scale apparent in North America or Japan. In Europe, MT systems are used mainly by large translation services and by multinational companies. For example, the software group SAP is translating some 8 million words a year using Metal and Logos, Ericsson is a large user of Logos for translating its manuals, and the European Commission has seen a rapid growth in the use of Systran, with demand now running at some 200,000 pages a year. Users are mainly non-linguist staff wanting translations for information purposes or drafts for writing documents in non-native languages.

In general, however, there continues to be widespread opposition among many professional translators to fully automatic systems. Instead, what they and most translation agencies and smaller companies prefer is the translator workstation approach.

Translator workstations

For professional translators, the attraction of the workstation is the integration of tools ranging from simple word processing aids (spelling and grammar checkers) to full automatic translation. The translator can choose to make use of whichever tool seems most appropriate for the task in hand. The vendors of these systems always stress that translators do not have to change their work patterns; the systems aim to increase productivity with translator-oriented tools which are easy to use and fully compatible with existing word processing systems. The four most widely used workstations all originate from Europe: Trados' Translation Workbench, IBM's Translation Manager, STAR's Transit,

Eurolang's Optimizer. In facilities and functions, each offer similar ranges: multilingual split-screen word processing; terminology recognition, retrieval and management; translation memory (pre-translation based on existing texts); alignment software for users to create their own bilingual text databases; retention of original text formatting; and support for a very wide range of European languages, both as source and target languages. Integration with MT systems is now provided by three of the workstations. In the case of Trados, access is provided to the Transcend software from Intergraph; IBM Translation Manager links up with Systran; and Eurolang Optimizer with Logos.

In addition, Europe has been the centre for most of the background and current research on workstations. The European Commission supported the major ESPRIT-funded TWB project (1989-94), involving 10 members from companies and universities, and it has supported TRANSLEARN, an LRE project for an interactive corpus-based translation drafting tool (a prototype translation memory system) based on EU regulations and directives from the CELEX (European Union law) database. A third project (DB-MAT), this time funded by Volkswagen, is investigating the use of a domain knowledge base integrated with a linguistic database as a translation tool; the languages are German and Bulgarian.

The Translation Service of the Commission itself is now developing its own workstation, EURAMIS. The aim is to optimise the efficiency of the translation resources already available, to create a database of translated EU documents (as a "translation memory"), and to provide easy access to MT systems. It will allow individual translators to develop their own tailor-made resources and facilities, with tools for text corpus management, glossary construction, and text alignment. Particular emphasis will be placed on the integration of MT and translation tools, including the mutual enrichment of Systran dictionaries and Eurodicautom lexical databases.

MT in Europe

Most of the cheaper PC-based MT software originates from Japan and the United States. In comparison, there have been surprisingly few MT systems developed and manufactured by European organisations.

Two come from the former Soviet Union: the successful Stylus system for Russian-English, English-Russian and German-Russian systems, and the PARS systems for Russian and Ukrainian to and from English.

Mention has been made already of the Langenscheidt T1 system, developed from Metal jointly with the Gesellschaft für Multilinguale Systeme (GMS), which succeeds Siemens as Metal agent for German versions (software and rights to the Dutch-French version are handled by the LANT company in Belgium). Also from Germany is the Personal Translator, a joint product from IBM and von Rheinbaben & Busch, based on the LMT (Logic-Programming based Machine Translation) transfer-based system under development since 1985. LMT itself is available as an MT component for the IBM Translation Manager. Both Langenscheidt T1 and the Personal Translator are intended primarily for the non-professional translator, competing therefore with Globalink and similar products.

Other PC-based systems from Europe include: Hypertrans for translating between Italian and English; the Al-Nakil system for Arabic, French and English; the Winger system for Danish-English, French-English and English-Spanish, now also marketed in North America; and the TranSmart system for Finnish-English from Kielikone Ltd.

As the methods and techniques of natural language processing become more familiar outside the research laboratories, many companies have been developing their own translation tools. Although a world-wide trend (e.g. PAHO and Smart in the US, and NHK, JICST and CSK in Japan), it is a distinctive feature of European MT activity.

Both Winger and TranSmart were initially built for specific customers. In the case of TranSmart, this was developed originally as a workstation for Nokia Telecommunications. Subsequently, versions were installed at other Finnish companies and the system is now being marketed more widely. A similar story applies to GSI-Erli. This large language engineering company developed an integrated in-house translation system combining an MT engine and various translation aids and tools on a common platform, AlethTrad. It has recently been making the system available in customised versions for outside clients.

Custom-built MT has become a speciality of Cap Volmac Lingware Services, a Dutch subsidiary of the Cap Gemini Sogeti Group. Over the years, this software company has constructed controlled-language systems for textile and insurance companies, mainly from Dutch to English. Probably the best known success story for custom-built MT is the PaTrans system developed for LingTech A/S to translate English patents into Danish. The system is based on methods and experience gained from the EUROTRA project funded by the European Commission.

The most distinctive feature of the European scene is the growth of companies providing software localisation, which are acquiring considerable experience in the use of translation aids and MT systems (e.g. Logos, Metal and XL8). A forum for the interchange of experience and the establishment of standards was set up in 1990: the Localisation Industry Standards Association; it publishes a newsletter (LISA Forum Newsletter) and produces a CD-ROM directory of products, standards and methods (the LISA Showcase). Ireland, as the main centre for these services, has its own Software Localisation Group and has recently set up a Localisation Resources Centre (with support from the Irish government and EU.)

MT research

As already indicated, nearly all the major computer software companies are showing interest in developing translation tools and systems, with Japanese and American companies in the vanguard; the growth in sales of PC-based products has revealed the huge potential market. By comparison, academic research has declined relatively in the area of written language MT as such (i.e. as distinct from speech translation and multilingual applications in language engineering).

Nevertheless, there is research on developing dialogue-based systems which combine computer-interactive authoring and translating into an unknown language, usually within a restricted subject field in order to ensure higher quality output. There is much interest in exploring new techniques in neural networks and parallel processing, and particularly in corpus-based approaches such as statistical text analysis (alignment, etc.), statistics-based generation from example texts, hybrid systems combining traditional linguistic rules and statistical methods, and so forth.

Above all, the crucial problem of lexicon acquisition (always a bottleneck for MT) is receiving major attention by many academic research groups, and the large lexical and text resources (such as those available from the Linguistic Data Consortium and ELRA) are being widely and fruitfully exploited. University MT research groups are increasingly working together with commercial organisations in order to develop customer-specific systems, e.g. Carnegie-Mellon University and the Caterpillar Corporation, or to undertake basic research for companies. However, the main emphasis of MT research has shifted to applications within the context of multilingual tools for specific needs, and to longer-term research on speech translation.

Since the ending of EUROTRA, research funds from the European Union have been more widely focused on projects within the broad field of language engineering, which includes multilingual tools of all kinds as well as translation assistance in various contexts. Practical implementation and collaboration with industrial partners is emphasised throughout, as well as the need for general-purpose and reusable products. Very many of these multilingual projects involve translation of some kind, usually within a restricted subject field and often in controlled conditions. It is not possible to describe here all those projects which involve multilinguality and translation. (For details see Jörg Schütz in MT News International No.15, October 1996.)

Speech translation

The goal of automatic speech translation was always a distant vision for MT, until developments in speech technology began to make it a feasible objective from the 1980s onwards. At first, research was on a small scale, e.g. the project at British Telecom to translate formulaic messages over the telephone. Later, in 1986, the joint government and industry company ATR was established near Osaka in Japan, and is now one of the main centres for automatic speech translation. The aim is to develop a speaker-independent real-time telephone translation system for Japanese to English and vice versa, initially for hotel reservation and conference registration transactions.

Other speech translation projects have been set up subsequently.

The JANUS system is a research project at Carnegie-Mellon University and at Karlsruhe in Germany. The researchers are collaborating with ATR in a consortium (C-STAR), each developing speech recognition and synthesis modules for their own languages (English, German and Japanese). In January 1993 the Consortium gave a successful public demonstration of telephone translation.

At Cambridge, SRI developed a speech recognition component for their Core Language Engine, and subsequently built a speech translation system between Swedish and English in the domain of air travel information. The SLT system has been operational since June 1993, and has also been adapted in recent years for speech translation to and from French and Spanish within the same domain.

In May 1993, the long-term VERBMOBIL project funded by the German Ministry for Research and Technology began. VERBMOBIL is intended to be a portable aid for business negotiations and to supplement users' own knowledge of the languages (German, Japanese and English).
Numerous German university groups are involved in fundamental research on dialogue linguistics, speech recognition and MT design; a prototype is nearing completion, and a demonstration product is targeted for early in the next century.

Speech translation is probably at present the most innovative area of computer-based translation research, and it is attracting the most funding and the most publicity. However, few experienced observers expect dramatic developments in this area in the near future. The development of MT has taken many years to reach the present stage: widespread practical use in multinational companies, a wide range of PC based products of variable quality and application, and growing use on networks and for electronic mail. Researchers know that there is still much to be done to improve quality.

MT associations

With the recent rapid developments, the establishment of MT associations to bring together users, developers and researchers has come at an opportune moment. The International Association for Machine Translation (IAMT) was founded in 1991 with three regional bodies: the European Association for Machine Translation (EAMT), the Asia-Pacific Association for Machine Translation (AAMT) and the Association for Machine Translation in the Americas (AMTA). Each of the three associations welcomes corporate members as well as individual members, and each maintains links to other related organisations in its region (e.g. EAMT has links with ELRA, EAFT, InfoTerm, etc.)

The IAMT has its own news magazine (free to members), MT News International, which carries information about new MT systems, new products, news of user applications, publications and conferences.

In early 1997, the IAMT will be publishing a Compendium of currently available MT systems containing details of languages, hardware and software platforms, user facilities, dictionaries and prices.

The IAMT holds a biennial conference, the MT Summit, which rotates between the three regions. In 1995 the fifth Summit was held in Luxembourg under the auspices of the European Commission. The sixth will be in San Diego, California, in October 1997, and will be organised by AMTA; in 1999, the Summit will convene in the Asian region, and in 2001 it will return to Europe. Each regional association also organises its own conferences and workshops: AMTA in 1994 (Columbia, Maryland) and 1996 (Montreal); EAMT in 1993 (Heidelberg), 1996 (Vienna) and 1997 (Copenhagen). All have successfully brought together developers, users and researchers to discuss common concerns and future prospects.

To join the EAMT, please contact:
EAMT Secretariat,
Attn: Christine Favre,
ISSCO,
54 route des Acacias,
CH-1227 Carouge (Geneva),
Switzerland
Fax: +41 22 300 1086.

For general information about the **EAMT,**
please e-mail eamt@cst.ku.dk.
John Hutchins
President, EAMT
University of East Anglia
Norwich NR4 7TJ
United Kingdom
E-mail: J.Hutchins@uea.ac.uk; or
100113.1257@compuserve.com

AVENTINUS: A multilingual information system for drug enforcement

Thomas Schneider

Most information systems are developed with the aim of increasing profits for the manufacturer. Others are developed to prevent the worst. One such project is AVENTINUS, a multilingual information system for drug enforcement, which is partially funded by the European Commission under the latter's Language Engineering program (LE1-2238).

The major reason behind the decision to fund the AVENTINUS project was the seriousness of the problem of organised crime, drug trafficking, money laundering and terrorism. It has become painfully clear that drug dealing poses one of the greatest threats to the nations of Europe today. The negative effects on society are plainly visible, with large sums having to be spent on health care for drug addicts. In addition, as drug addiction usually impairs the ability of users to hold down normal jobs, unemployment benefits or social security payments must be made. What is more, the need to obtain expensive drugs leads to criminal behaviour ranging from prostitution to theft, robbery and even murder - all of which affect people outside the drug scene.

At the same time, economic constraints prevent countries enlarging their police forces enough to be an effective countermeasure. Established international networks of drug dealers are becoming a powerful economic factor, influencing a whole range of activities in all sectors and creating a shadow economy which is not controlled by state authorities. In short, the effect on the European economies cannot be underestimated, and if no means can be found to combat drug trafficking effectively, follow-up costs will be staggering. The damage attributable to drug dealing in Germany alone is estimated at around ECU 5 billion per year, and is growing exponentially. Extrapolate this local damage to a global scale, and the figures are staggering.

It is self-evident that drug enforcement must be internationally co-ordinated. Go-it-alone national efforts are likely to fail since criminal networks operate on an international scale (a problem which has increased since the break-up of the Soviet Union and the opening of the EU's eastern borders).

Although European police agencies and government institutions have been co-operating with each other, major obstacles to efficient information exchange and rapid response currently exist. First of all, although relevant information may available, officers may not know where to look. In addition, this information may be spread over several different sources, e.g. chemical descriptions of substances, photographs, police reports, person files, or video sequences. What is more, it resides on different computers in different countries - and in different languages, not all of which will be spoken by the person requiring it. What matter if today we can transmit data packages within fractions of a second - if the content is incomprehensible, the speed is irrelevant.

Within the AVENTINUS project, the members of the user group - a number of law enforcement agencies - are responsible for defining requirements, since only they have the background to decide which technical solutions can contribute to their practical work. In general, there are two scenarios which need to be supported: The first one involves the acquisition and interpretation of data, and the second the retrieval of information which is relevant to the case in hand.

Data acquisition and interpretation

Each of the government agencies in the project receives numerous items of information every day: as plain text, formatted messages, telexes, faxes, or video sequences, all from a wide range of different sources and in different languages. In current workflows, the first step (which is not within the remit of the AVENTINUS project) is to convert any paper documents into machine-readable text. Next, an attempt is made to identify the language involved - not always a trivial task if your expertise is not in Asian or Eastern European languages. After this, the text has to be translated by a translator skilled in the relevant language, and the content described and stored in a database for future use. This is a lengthy process which takes not only a lot of time but also considerable resources.

Within AVENTINUS, the process will be automated as far as possible. A computer program will automatically identify the language involved and the document will then be channelled to a machine translation system, if one is available for the language in question. Alternatively, the document will be passed by a translation memory or, if this also proves ineffective, a terminology database. In the latter case, terms detected in the database are inserted into the original text. While this cannot replace a translation, it can give analysts at least a hint as to the subject matter. This can be especially helpful if the text is written in a language for which no machine translation system is available. Such an assessment of a text's relevance makes it easier to decide if it warrants translation, or if it is less urgent.

Effective information retrieval

The second scenario involves the retrieval of information which is relevant to the case in hand. Agents in drug enforcement units need quick access to information, be it on a person's movements, on the properties of a certain drug, or on current narcotics legislation. Usually this information is available somewhere, but retrieval of the relevant facts is time-consuming. Conventional techniques for querying full-

text databases produce more "noise" than accurate hits, and in the past officers had to be proficient in the language of the source. In addition, in a large database of perhaps millions of text passages, a search based on key words is likely to produce more answers than can be looked at in detail. On the other hand, the same concept might be expressed in a different surface form, with the relevant concept described perhaps in a subordinate clause or expressed by a synonym. In such cases, it probably will not be found at all, or a user would have to know all possible expressions in advance.

In AVENTINUS, linguistic analysis of the texts is coupled with the automatic generation of conceptual links between content words on the basis of a statistical analysis of their contextual occurrence. This method produces a much higher percentage of correct and focused answers. A fuzzy search module will also allow misspelt words or names which have been transliterated in different ways (for example, Cyrillic names are usually represented differently in German and English) to be found.

In the past , in order to query foreign language databases, users had to express their questions in that language, once again trying to anticipate different expressions for single concepts - a time-consuming and error-prone procedure. By contrast, AVENTINUS will enables officers to formulate queries in their native languages, with the queries then being translated into the appropriate form and the appropriate language for the various databases in a manner which is totally transparent to users. Some data such as chemical tables are stored in structured for, while others are hidden in running text. By using sophisticated syntactic and lexical analysis and building dependency relations between concepts, AVENTINUS will either retrieve the relevant documents and avoid overloading users with digital garbage, or extract the relevant facts from the texts stored. Fact extraction is a highly complex problem, as it requires the development of intricate but still manageable domain models, hyperlinks between concepts and access to databases with e.g. named entities. Machine translation, translation memory and term substitution systems will ensure that the user receives the requested information in his own native language. In other words, AVENTINUS will combine information from both structured and unstructured databases into an understandable package, even if the sources reside in different countries and are formulated in different languages. This should speed up searches for relevant information and greatly improve accuracy and coverage.

Taking the mountain to Mohammed

Two aspects had absolute priority during the project structuring phase. Firstly, in contrast to most research projects, AVENTINUS is strictly user-driven. Not only do all technical specifications have to be approved by the user group, but any newly developed or existing tool must fit into their existing infrastructure. No agency can afford to scrap in-house IT systems which have been developed and filled with data over many years in favour of a radically different approach. In some cases, using newer technologies would provide efficiency gains, but their introduction would severely disrupt internal workflows and established procedures. Academic research and long-term goals thus have to be subordinated to practicality.

Secondly, AVENTINUS cannot and does not want to reinvent the wheel. If components are available on the market which fit into the overall structure, they will be used. It is a fact of life that newly developed complex systems invariably suffer from a multitude of bugs, many of which will only be found after extended periods of use in real-life situations. In the case of the planned applications, such risks cannot be taken. However, together, the AVENTINUS Consortium partners have more than a thousand person-years of experience in the area of natural language processing in product development environments as well as research. The main tasks to be addressed are the integration of the various components, the development of additional language-specific modules such as lemmatizers, and the building up of large amounts of data such as multilingual lexical entries. A high proportion of overall project resources is earmarked for testing.

Project planning and participants

The AVENTINUS project was divided into two distinct phases. The first one, scheduled to last sixteen months and now completed, collected and evaluated user their expertise are the Bundeskriminalamt (BKA) and the Amt für Auslandsfragen (AFA) in Germany, the Spanish CESID and the Swedish Rikspolisstyrelsen.

The AVENTINUS Consortium is co-ordinated by Gesellschaft für multilinguale Systeme (GMS), an industrial company with extensive experience in machine translation, information retrieval, networking and systems integration. The other development partners, INCYTA of Spain, ILSP of Greece, and the Universities of Sheffield and Gothenburg, are all specialists in natural language processing. Since all partners have co-operated well in the past and since AVENTINUS has been structured as a highly pragmatic project, the chances are good that by the end of the second phase (20 months) there will be a robust, operative system with a wide spectrum of potential applications.

Due to financial limitations, AVENTINUS cannot address all domains and all European languages, let alone interesting but "expensive" languages such as Arabic, Chinese, or Farsi. However, if the implementation phase is as successful as it currently looks, it is hoped to extend the project to other languages. If the operative system manages to reduce the damage inflicted on us all by drug pushing by only 2%, the project would be cost-effective even if it meant investing 10 billion ECU.

For further information please contact:
Gesellschaft für multilinguale Systeme
Balanstr. 57
D-81541 Munich
Tel.: +49-89-49042031
Fax: +49-89-49042020
E-mail: info@gmsmuc.de
http://www.gmsmuc.de

The Amaryllis Project - Access in French to textual information

Annie Coret

Introduction

The Amaryllis project was initiated and is sponsored by Aupelf-Uref (Agence francophone pour l'enseignement supérieur et la recherche - the Agency for Higher Education and Research in French-speaking Countries) and DISTNB (Direction de l'Information Scientifique, des Nouvelles Technologies et des Bibliothèques - the French Ministry of Education, Higher Education and Research 's Directorate of Scientific Information, New technology and Libraries). In July 1994, the former launched a call for tenders (ARC A1) to encourage the development of systems for access in French to textual information by creating a corpus of documents, queries and answers similar to those created in the United States in the TREC (Text REtrieval Conference) project. (In the latter, large, identical corpora of documents in English and search topics are supplemented by the expected answers in order to facilitate training in the use of the systems, after which new corpora and topics without any predefined answers are used to test the systems. The results obtained are then compared using an identical methodology and discussed in a joint meeting) DISTNB's goal in the project was to promote the systems and encourage greater market awareness.

Amaryllis is divided into an exploratory phase, due to end in April 1997, in which preliminary tests are being carried out on the basis of an existing methodology, and a second phase, in which the experience of the first phase would be taken into account. Three different types of organisation are involved: the co-ordinator (INIST-CNRS[1]), corpus suppliers selected and sponsored by Aupelf-Uref (INIST-CNRS, LRSA[2] and OFIL[3]) and test participants who are not financed, but who answered a call for participants launched by the organiser.

Corpora

Each supplier is compiling three types of corpus: documents in French, search topics formulated in French and retrieved "hits". The use of the LRSA corpus (books on Melanesia) has not been possible in the exploratory phase, as an agreement had to be concluded with the book publishers. The OFIL documents comprise the title and text of articles from the daily newspaper Le Monde, with each set of OFIL documents containing three months' work. The INIST documents consist of titles and abstracts from the PASCAL and FRANCIS bibliographic databases covering all types of domain: science, technology, medicine, social sciences, etc.

The search topics were created by information specialists on the basis of actual questions submitted by journalists (OFIL) and end users (INIST). In principle, they include all the information necessary to understand the problem to which they relate, and to judge the relevance of the answers.

The "Domaine" field allows a knowledge field to be assigned to the topic, while the "Sujet" defines the topic. "Question" corresponds to the statements of user needs, "Complément" provides information on the documents to be extracted, and "Concepts" contains a set of descriptors for the retrieval field. The answer files were compiled by the different supplier using their own tools and methods.

Two CD-ROMs have been produced during the exploratory phase.

CD-ROM 1 comprises
- an OFIL corpus: 11,000 documents (34 MB), 11 topics and their answers
- an INIST corpus: 163,000 documents (68 MB), 15 topics and their answers

CD-ROM 2 comprises :
- an OFIL corpus: 10,500 documents (34 MB), 15 topics
- an INIST corpus: 151,000documents (64.5 MB), 15 topics.

The documents have been formatted using SGML and ISOlatin character coding. They may only be used by the participating teams within the context of the project.

Progress so far

The *exploratory phase* is based on TREC and therefore on an existing methodology, although it remains experimental in nature. It differs from TREC chiefly in its scale (the volume of data is markedly lower), in the separate treatment of the OFIL and INIST corpora, and in the method used to compile answer files (by suppliers at the same time as the tests are performed by participants).

The participants use the data on CD-ROM 1 to learn and to prepare the system for the evaluation tests, after which they are not allowed to modify their system. They then use the data on CD-ROM 2 to perform two types of evaluation tests:
- tests simulating direct routing : the formulations of topics created during the training phase are "applied" without modification to the new batches of documents supplied, and
- tests simulating a search : the new topics supplied are "applied" to the known documents.

Suppliers revise their respective answer files in the light of the participants' results and create a reference file of answers. The results from each system are then evaluated using the TREC analysis software. This gives the precision (i.e. the number of relevant documents found compared to the total number of documents) and the recall (i.e. the number of relevant documents found compared to the total number of relevant documents) of a batch of answers from a test participant for each topic or group of topics, and compares them to the reference answers compiled by the suppliers.

Conclusion and outlook

The exploratory phase will produce an analysis of the work to be carried out and the scientific, technical, organisational and legal difficulties to be resolved for by the co-ordinator, the corpus suppliers and the test participants. An in-depth progress report will be produced before the next phase is begun.
References are available on our Web site.

1 Institut de l'Information Scientifique et Technique (Institute for Scientific and Technical Information)/Centre National de la Recherche Scientifique (National Center for Scientific Research)
2 Laboratoire de recherches sémiographiques en anthropologie (Laboratory for Semiographic Research in Anthropology)
3 Observatoire Français et International de Industries de la Langue (French and International Observatory for Language Industries).

1996 ELRA General Assembly

The following report provides a summary of the second ELRA General Assembly meeting, held in Le Louvre, Paris on 20 December 1996. The full minutes will, of course, be sent by separate mail to all ELRA members.

Following a welcome by the Chairman of the meeting, Professor Antonio Zampolli, the number of members present (24) and proxies (10) was established and the agenda approved as distributed. The minutes of the 1995 General Assembly were adopted without modification, and the Assembly told of the results of the Board meeting held immediately before the General Assembly, in which all Board members had been re-elected unanimously to their positions.

Professor Zampolli then briefly outlined the progress ELRA had made in 1996, including the setting up of the Central Distribution Unit (ELDA), the collection of a considerable number of language resources, the production of the first catalogue and the creation of practical distribution agreements.

These achievements were echoed by Dr Khalid Choukri, the CEO of ELRA, who reported in detail on the period from October 1995 to September 1996. His first task had been to set up the Distribution Agency (ELDA), which was now operational and, despite some initial difficulty in finding suitable candidates, now had a staff of 4 people. He also outlined the various panels of experts, their role and their convenors. The total number of ELRA members was now seventy, seven of whom had applied for membership in 1996. Membership is balanced across the three colleges but not across Europe, and membership drives will focus in future on poorly represented countries. The first resources catalogue had been very unbalanced in its coverage of the three colleges, but the second release showed substantial improvement. Dr. Choukri then outlined the logistics problems which had been addressed with regard to distribution, including the legal issues touched on in the drawing up of distribution/VAR/end-user agreements, and the publication of the Guide to Terminology Agreements. He also described a number of links established with other organisations.

All deliverables under the projects funded by DG XIII (LE1-1019 (ELRA) and LRE-62050 (MULTEXT)) had been submitted on time. Further funding had also been obtained, mainly from DG XIII (LE Programme) and the French Government. Dr. Choukri expressed his regret that other national European governments did not contribute to the ELRA budget.

Publicity activities included regular publication of a bilingual newsletter (English & French), the ELRA Web site (bilingual access in French and English, plus additional services reserved for members), and revised marketing brochures.

Vice-president Dr. Joseph Mariani (LIMSI CNRS) then reported on the work of the Panel of Experts for Distribution and Pricing Policy. He defined ELRA's strategy as getting members, getting resources and selling resources, and gave examples of prices and of the substantial discounts offered to members. Comparing ELRA with its counterpart in the USA, LDC, he stressed that ELRA had less money and fewer staff, but more members during the first year. It also had different pricing and membership policies, but the same range of prices. He reported on a proposal that all LDC members could become ELRA members or subscribers direct, while ELRA members would pay only the difference in membership fees. He asked for comments and feedback on new means of distribution (electronic networks) and on the co-operation schemes described.

Dr Khalid Choukri then defined his main goals for 1997 as supplying more services to members, improving the quality of the catalogue (in term of the descriptions and samples given, etc.), and active marketing. A more technical issue is validation — a key factor in distribution, and one that adds value to resources. It will be achieved via strong links with the relevant LE projects (SpeechDat for speech resources, and INTERVAL for terminology) and via a subcontract for written resources. Board member Bente Maegaard (Center for Sprogteknologi) then explained the issues behind the validation of written resources (particularly lexicons) and the different types of validation involved (formal validation, content validation, compliance with standards, purpose of validation, etc.). What was needed was a validation methodology laid down in a workable manual, and a set of sites where validation could be performed.

Dr. Thomas Schneider, ELRA Treasurer, presented the Financial Report for the fiscal year October 1995 - September 1996). The positive state of the accounts was partly due to the unexpected delay in setting up ELDA and the consequent postponement of staff costs. Following his overview of major income and expenditure for the year under report, he underlined that income from resources will remain poor in 1997, and drew attention to the additional expenditure required (e.g. for hiring a marketing specialist). In response to members' suggestions, it was decided to distribute a detailed financial report together with agenda for future General Assemblies, along with a short- and medium-term business plan.

The Management Report and the Financial Report for 1996 and the budget plan and activities for 1997 were formally approved by the General Assembly.

After some further discussion, Professor Antonio Zampolli closed the meeting at c. 17.00 and thanked the participants for coming.

EAGLES: A brief progress report

John McNaught, EAGLES Co-Chief Editor

The Expert Advisory Group on Language Engineering Standards (EAGLES) stems from an initiative of the European Commission. It was launched in February 1993 within the Linguistic Research and Engineering programme, as project LRE-61-100. The aim of EAGLES is to accelerate the provision of standards for: very large-scale language resources (such as text corpora, computational lexicons and speech corpora); the means of manipulating such knowledge via computational linguistic formalisms, mark up languages and various software tools; and the means of assessing and evaluating resources, tools and products.

The LE-EAGLES project (LE-34244), beginning now in 1997 and seen as a continuation of LRE-EAGLES, aims to pursue the same general objectives over the 20 months of its existence. On the one hand, this involves eliciting feedback on late-stage LRE-EAGLES results, incorporating feedback and disseminating the revised material widely and, on the other hand, further work will be carried out to identify, deepen and broaden certain aspects of language engineering deemed mature enough for a standardisation push. The work of LE-EAGLES is thus to be seen within a long-term standardisation initiative, applying an already proven methodology to relevant areas of language engineering and to elicit feedback and disseminate revised and new material on a cyclical basis.

The motivation for EAGLES

Progress in NLP and speech applications is hampered by a lack of generic technologies and of reusable language resources, by a proliferation of different information formats, by the variable linguistic specificity of existing information and by the high cost of developing resources. In general, it has never been possible to build on the results of past work, whether in terms of resources or the systems that use them.

The whole field of language engineering is dependent on at least de facto standardisation to allow further development and particularly greater progress in language engineering applications. Without language engineering standards underlying language engineering applications and resources, users of language technology will remain ill-served. The application area will continue to be severely hampered and will continue to see success only in niche or highly specialised applications (e.g. speech aids for the disabled; spelling checkers).

The potential impact of language engineering standardisation is thus significant, in that it will open up the application field, allow expansion of activities, sharing of expensive resources, reuse of components, rapid construction of integrated, robust, multilingual language processing environments for end-users, etc. Language processing, for the end user, will become second nature rather than an obstacle course as at present. With language engineering companies in the EU producing low-cost, effective applications based on language engineering standards developed for their purposes, the potential for such companies to become world market leaders is clear.

Since the formation of EAGLES, work in the EU on language engineering standards has largely been concentrated within this initiative. Related efforts elsewhere are closely linked with EAGLES and feed off it. Besides developing from pre-existing preparatory work on the lexicon, on corpora and on speech aspects, EAGLES has initiated standards-related work in the areas of NLP formalisms and in the evaluation of NLP systems. Evaluation in particular has proved to be a point of focus for many interested groups and projects throughout the world and the EAGLES evaluation group has had significant input from them. The same has happened to the Lexicon and Corpus groups, whose recommendations are already applied in a large number of European and national projects. Indeed, EAGLES has acted as a catalyst and testing-ground.

Several LRE projects have been active in contributing comments and in testing EAGLES proposals, thus offering a concrete industry-related setting. Given the amount of industrial participation in EAGLES itself, it is notable that there has been significant advance in language engineering standards over the past two years, thus re-emphasising the need to involve industry in such efforts in targeting clearly identified and motivated standardisation goals.

The EAGLES results to date are to be seen as a first step on the path towards standardisation for language engineering purposes. The immediate future is thus the time for consolidation, testing, refinement and dissemination of these initial results, via a second cycle of EAGLES-like activity, as well as for the development of proposals covering new areas.

LE-EAGLES is firmly positioned as the natural continuation of the work begun in LRE-EAGLES, which has been positively received, in the inherently long-term standardisation process. LE-EAGLES, as LRE-EAGLES, brings together representatives of major collaborative European R&D projects in relevant areas, to determine which aspects of our field are open to short-term de facto standardisation and to encourage the development of such standards for the benefit of consumers and producers of language technology. This work is being conducted with a view to providing the foundation for any future recommendations for International Standards that may be formulated under the aegis of ISO. EAGLES' work towards de facto standards is intended to allow the field to establish broad consensus on key issues, thus providing an opportunity for consolidation and a basis for technological advance and the expansion of knowledge.

Dissemination

Dissemination lies at the heart of standardisation activities, as much to gain feedback as to encourage uptake of the results. The main results of EAGLES currently take the form of handbooks of best practice and guidelines. Every advantage is being taken of new technology and new media in dissemination. The existing LRE-EAGLES initial guidelines are being converted for the World Wide Web (http://www.ilc.pi.cnr.it/EAGLES/home.html), as well as being published in traditional form. This will ensure very wide dissemination and, via WWW, a straightforward path for instant access to updated or new guidelines. Dissemination via CD-ROM, FTP, etc., is also appropriate. The final handbooks are expected to be a set of fully hyperlinked documents primarily designed for dissemination via the Web.

Partners

The LE-EAGLES partners are: CPR (Italy), Instituto Cervantes (Spain), RXRC (France), GSI-Erli (France), University of Lancaster (United Kingdom), Universität Bielefeld (Germany), Universität Stuttgart (Germany), Vocalis Ltd. (United Kingdom), DRA (United Kingdom), CST (Denmark), Sharp Laboratories of Europe (United Kingdom) and ISSCO (Switzerland). It is to be noted, however, that a substantial part of the manpower required to accomplish the tasks will be provided on a voluntary basis by the parti-

cipating organisations or by individual experts. This was also the case in LRE-EAGLES (in which some 200 organisations were latterly involved).

Themes

As a result of the favourable LRE-EAGLES final review and discussions with the European Commission, a number of novel features or changes with respect to the previous set-up have been introduced for LE-EAGLES. In particular, the themes of the project have been extended to encompass work on (1) lexicon encoding, including semantic encoding and labelling, encoding of multiword expressions, and bi- and multilingual aspects in lexicon encoding; (2) fostering the integration of spoken and written language resources; and (3) revisions and completion of the spoken language and evaluation handbooks. It is to be noted that the themes under (1) can only now be considered: the degree of consensus over enabling levels of linguistic information is now well advanced and thus these new themes become viable for de facto standardisation. The work on evaluation will also follow a slightly different path to that on the other main themes, in that it will involve the organisation of workshops in co-operation with industry to tackle aspects of language engineering evaluation with an aim to developing further de facto standards.

Co-operation

Right from the beginning, EAGLES has been active in establishing links with related bodies and associations. This activity will increase in LE EAGLES. In particular, there are agreements with ELRA (to co-operate in the continuous dissemination of EAGLES results; to organise a User Forum for the discussion of EAGLES recommendations; to promote adoption of EAGLES recommendations and guidelines; and to base ELRA validation activity on EAGLES recommendations), ELSNET (to assist in the dissemination of EAGLES information and results, in particular through articles and news regularly published in the ELSNews bulletin; and to validate, via the ELSNET Resources Task Force, the recommendations of the EAGLES Spoken and NLP integrated Language Resources Working Group), various language engineering projects engaged on building language resources and various COPERNICUS projects (with a view to defining and validating guidelines for Eastern European languages).

Further details of EAGLES may be obtained from:

EAGLES Secretariat
ILC-CNR
via della Faggiola 32
Pisa
I-56126

Tel: +39.50.560481
Fax: +39.50.556285
E-mail: eagles@ilc.pi.cnr.it
http://www.ilc.pi.cnr.it/EAGLES/home.html

LE sector events

ACL'97/EACL'97 Joint Conference

The 35th annual meeting of the Association for Computational Linguistics and eighth conference of the European Chapter of the Association for Computational Linguistics will take place on July 7-12, 1997, at the Universidad Nacional de Educación a Distancia (UNED) in Madrid (Spain). Several workshops are being planned as satellite events, including ones entitled "From Research to Commercial Applications", "Making NLP Technology Work in Practice", "Intelligent Scalable Text Summarisation", and "Natural Language Processing for Communication Aids".

First JST FRANCIL 1997

The first JST FRANCIL (Journées Scientifiques et Techniques) will be held in Avignon, France on April 15-16, 1997. The aim of the conference, the title of which is "Language Engineering - from Research to Product", is to chart the progress made in linguistic engineering. Ten main topics will be covered: speech recognition, speech understanding, speech synthesis, information retrieval from document databases, multilingual text alignment, automatic extraction from terminological databases, text understanding, spoken and written language learning systems, the evaluation of linguistic engineering systems and linguistic resources for automatic language processing.

LISA Forum and Workshop

The next Forum organised by LISA, the Localisation Industry Standards Association will be held in Mainz, Germany on 5-7 March, 1997. The theme is "How Technology solves Localization Business Problems". Key topics include implementing language processing tools, European support for multiple language software and documentation, legal requirements for companies selling on the European market, multilingual support for the Web, Unicode, terminology interchange and the current trend towards consolidation in the localization industry. A separate workshop on 4-5 March, being hosted jointly by LISA and Fry & Bonthrone, will take an in-depth look at multilingual information management. Current and future user requirements and expectations will be presented, as will ongoing projects and tools, and analyses of implementation scenarios and design specs for future tools and procedures.

For more information about **ACL'97/EACL'97** access

http://horacio.ieec.uned.es/cl97

For more information, please contact:
JST'97 FRANCIL
LIMSI-CNRS
Tel: +33 1 69.85.80.80
Fax: +33 1 69 85 80 88
E-mail:jst97@limsi.fr
http://www.limsi.fr/Recherche/FRANCIL/JST97.html

For more information on both events, please contact the **LISA Secretariat** at:
2 bis rue Ad Fontanel, CH-1227 Carouge/GE, Switzerland.
Tel: +41 22 301 5760
Fax: +41 22 301 5761
E-mail: lis@lisa.unige.ch
http//.www.lisa.unige.ch

Computer software for automatic knowledge base expansion

Kenji Sugiyama, EDR

This article describes the new 2-year project which EDR launched in its 1996 business year with funding from the IPA (Information Technology Promotion Agency). The research is being conducted in co-operation with Prof. Hozumi Tanaka and Prof. Takenobu Tokunaga (Tokyo Institute of Technology) and Prof. Junich Tsujii (Tokyo University).

The aim of the new project is to develop the core of an algorithm/software for automatically accumulating and building up different types of knowledge (such as technical trends, industry trends, transportation information, etc.) needed by research organisations, companies and individuals; to support the creation of such knowledge bases and the development of application software by promoting the use of this core, and ultimately to support companies and individuals in their daily work.

The various types of knowledge described above are often expressed using languages - typical examples are newspaper and magazines. The project therefore builds on linguistic knowledge to develop software which automatically builds up a linguistic knowledge base. The input for this software is a document such as a newspaper, and the output is the knowledge of grammar, vocabulary, and concepts (i.e. word meanings).

Multimedia information such as documents are currently being digitised on a large scale and distributed via computer networks (the Internet is the typical example here). As a result, we are about to experience significant social change, as evinced by such new concepts as electronic commerce and the virtual organisation. In such an environment, technology for automatically extracting linguistic knowledge from widely circulated electronic documents has an important contribution to make to the future information society. It can be used as a retrieval system to find valuable information which would otherwise remain buried in the flood of data, as a decision support system for more effective business management and research activities, or as a core technology for concurrent engineering in order to support more efficient development and production activities.

Scope, structure and implementation

The following topics must be addressed in order to achieve these objectives:

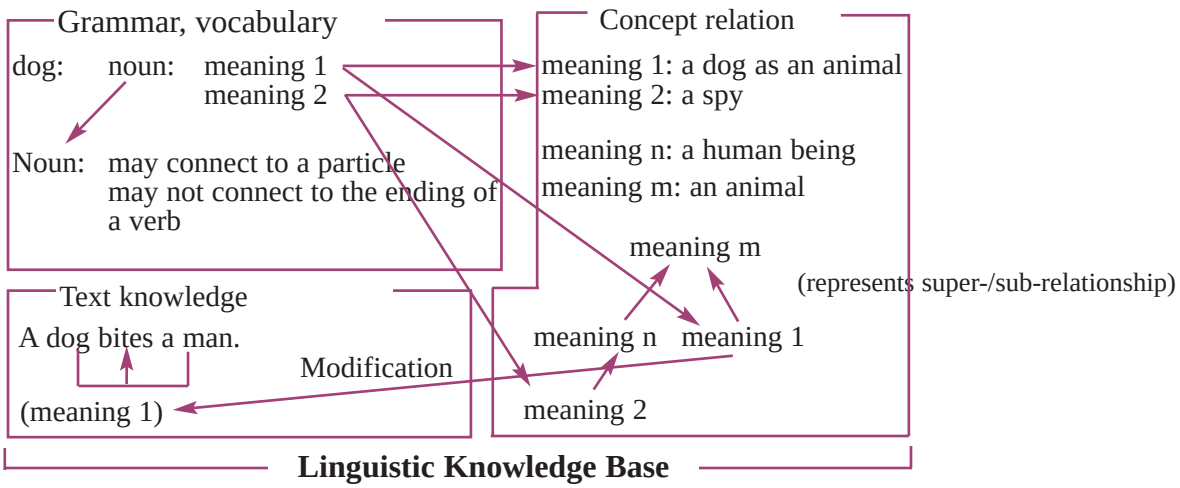
Linguistic Knowledge Extraction Software

- Morphological, syntactic, and semantic analysis
- Knowledge extraction and consistency

Knowledge Base Management Software

- Knowledge retrieval
- Knowledge registration
- Knowledge update

Relationships within knowledge base (image)



The input for the linguistic knowledge extraction software is provided by documents such as newspapers, books and technical documents, while the output is the knowledge of the words, grammar, concept relations and texts. The processing is done by the sub-modules. The morphological, syntactic and semantic analysis submodule extracts individual words from a continuous text string, finds modificational relationships between words, and extracts sentence meanings. The knowledge extraction and consistency submodule extracts knowledge such as words and meanings from the results of the analyses and ensures that they are consistent with the knowledge which has already been accumulated. The knowledge stored in the knowledge base is used during processing as needed. The linguistic knowledge output by the linguistic knowledge base extraction software is accumulated in separate knowledge bases (a grammar and vocabulary knowledge base, a concept relation knowledge base and a text knowledge base) by the knowledge base management software.

The research on and development of the linguistic knowledge extraction software is based on the theory and methodology of the MSLR parser (a morphological and syntactic analysis tool) developed by the Tokyo Institute of Technology, which will be enhanced and extended during this project.

As regards morphological and syntactic analysis and the extraction of grammar and vocabulary, we will develop a morphological-level (word-level) grammar known as a "morphological connection table", Japanese grammar rules (probabilistic context-free grammar) and a prototype to extract vocabulary such as proper nouns. This will be done in the first year of the project. For semantic analysis and concept relation knowledge extraction, we will design and develop a software which extracts the concept relations (semantic relations) between words or basic words, together with a method of analysing statistically what other terms may appear in the context of the term in question and organising them into clusters (semantic clustering), and of analysing the relationship between basic words within the terms selected (semantic relations). The results will be evaluated and improved on in the second year in order to obtain the final results.

The most important issue in this knowledge extraction process is to determine the kind of information from which new knowledge should be found. Currently, we use a comparison of information with existing knowledge (i.e. inconsistencies with existing knowledge or new information which is not stored in the knowledge base) as the trigger. In addition, we shall be performing analyses and tests to determine what should trigger knowledge extraction. In order to ensure efficient testing, evaluation and improvement, we shall develop a tool to display the extracted knowledge visually, in an easily understood manner.

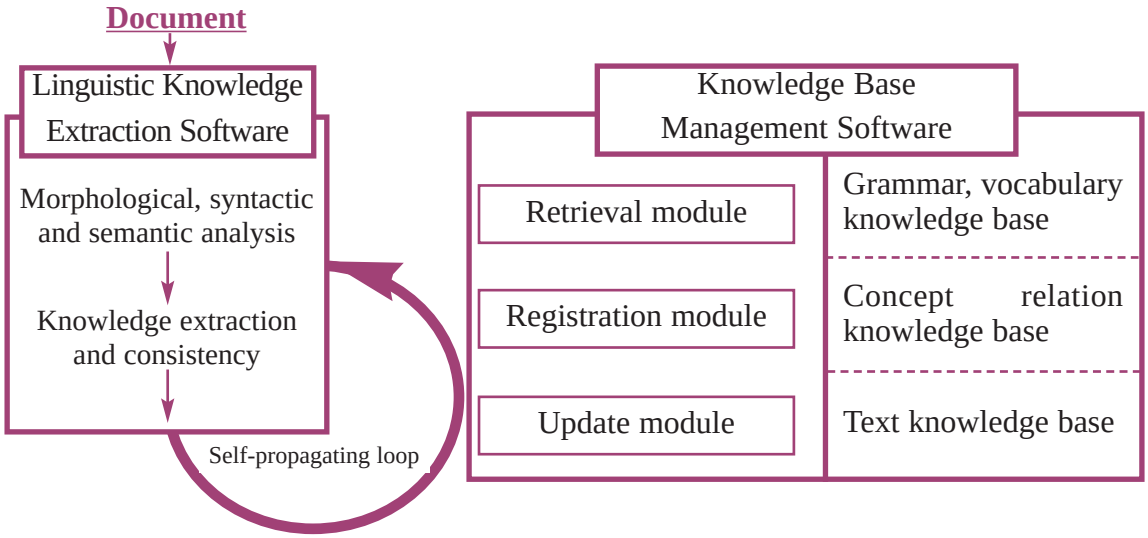
With respect to the knowledge base management software, we will develop flexible knowledge retrieval, registration and update sub-modules to allow knowledge consistency to be easily maintained. In this context, we will build on our experience in developing a large scale electronic dictionary management system for EDR, and use the technology to ensure consistency of dictionary management.

The EDR Electronic Dictionary will provide the initial data for the linguistic knowledge base (grammar and vocabulary knowledge base, concept relation knowledge base and text knowledge base); it will then be enhanced in a self-propagating manner using the linguistic knowledge extraction software mentioned above.

EDR Electronic Dictionary, which was released in 1995, is composed of ten subdictionaries: a Japanese Word Dictionary, an English Word Dictionary, Japanese-English and English-Japanese Bilingual Dictionaries, a Concept Dictionary, Japanese and English Co-occurrence Dictionaries, a Technical Terminology Dictionary, and the EDR Japanese and English corpora.

Currently, 22 private companies and 84 universities in Japan use the EDR Electronic Dictionary, along with five universities

For more details, please access the EDR Web site:
http://www.iijnet.or.jp/edr/
Kenji Sugiyama
Japan Electronic Dictionary Research Institute, LTD. (EDR)
78-1, Kanda-sakumagashi, Chiyoda-ku, Tokyo 101, Japan
E-mail: kenji@edr.co.jp



Implementation structure

New Resources

ELRA-M11 German-English lexicons (GMS)

METAL famous lexicons especially designed for automatic translation. The lexicons contain bilingual entries including the indication of word class and subject area. Several domains are available:

Agriculture (1,872 entries),
Common Administrative (2,742 entries),
Common Technical Vocabulary (18,031 entries),
Data Processing, including Hardware, Software, Data Transmission (44,484 entries),
Economics and Finance (4,641 entries),
Electrical Engineering (16,689 entries),
General Vocabulary (54,037 entries),
Mechanical Engineering (5,820 entries),
Biology and Chemistry (9,405 entries),
Geology (4,409 entries),
Physics and Mathematics (2,627 entries),
Automobile Technology (1,149 entries),
Optics (3,775 entries),
Shipbuilding (1,636 entries),
Textile (5,354 entries),
Typography (2,123 entries),
Telecommunications (24,395 entries).

ELRA-T104 VERBA (Ediciones Verba) Polytechnical and Plurilingual Terminological Database

Entries for English-Spanish:

Scientific research & mathematical sciences (906 entries), Geosciences (10,215), Computer science, electronics & telecommunications (70,580), Industry (47,578), Transport & Maintenance (12,291), Economy (145,572), Biological sciences (38,989), Communication & media (8,143), Chemical & physical sciences (27,467).

Entries for English-French-German-Spanish:

Environment (36,658), Health (66,727), Agriculture & food (25,975), Construction & public works (8,429), Law & policy (56,578), Sports & Leisure (17,312)

Two specialized lexicons:

Spanish-English and English-French-German without domain codes: electronics, telematics, law, taxes, customs, etc. (550,000 entries).

Two general lexicons:

Spanish-English-French-German and Spanish-English-French-German-Portuguese-Italian (83,000 entries).

This terminological database contains, for each domain, a sub-domain indication is given (from 2 sub-domains for Scientific research to 39 for Sports & leisure). Each entry consists of a definition, phraseological unit, abbreviation, usage information, grammatical labels.

Format: ASCII
Medium: Floppy disk

ELRA-W14 Monolingual Greek corpus (ILSP -Institute for Language and Speech Processing)

Monolingual Greek corpus of 1 million words. The corpus consists of articles written in 1996 from the Greek daily newspaper ELEFTHEROTIPIA. Each file contains annotated text with SGML mark-up accompanied by a text header.

ELRA-M12 English-German lexicons (GMS)

METAL famous lexicons especially designed for automatic translation. The lexicons contain bilingual entries including the indication of word class and subject area.The following domains are available: Common Technical Vocabulary (11,300 entries), Data Processing (5,524 entries), General Vocabulary (60,736 entries).

ELRA-L17 « N de N » Dictionary (compound nouns) (CORA)

This dictionary contains 21,000 compound nouns of un inflected « N de N » groups, classified in 1,000 human entries (divided into job, group, animated), 4,200 concrete entries (divided into clothes, dishes, furniture...), 6,000 abstract entries (divided into tables of auxiliary verbs such as : « avoir », « donner »,etc.), plus syntactic/semantic information about determiners, verbs, etc.

Language: French
Format : Tagged ASCII
Medium: Floppy disk

ELRA-T102 Idioms, Proverbs and General Expressions

This dictionary gathers all figurative expressions, locutions, proverbs, idioms etc., in French as well as in English with their translations and opposite. This issue includes 5,000 pairs of English-French expressions. Their synonyms, where applicable, are mentioned in a field of the record. Significant words of the expressions are located in a specific field. Expressions are in the infinitive form or conjugated.

ELRA-L19 English lexicon (CORA)

Entries: 160,000.
Language: English
Format: ASCII
Medium: Floppy disk

The dictionary is divided into 4 main syntactic categories: nouns (93,500), verbs (35,800), adjectives (46,600), grammatical words (8,865). The lexicon contains a list of inflected words with corresponding syntactic categories and lemmas. Each entry is tagged with specific separators. A single word corresponds to a single entry in the lexicon.

ELRA-L16 Tri-, quadri-, pentagrams dictionaries (CORA)

Sequences: 5,487
Format : ASCII
Medium: Floppy disk

The dictionaries consist of a list of sequences of 3, 4 or 5 characters which follow each other in French words. In particular, they enable users to locate misspelt sequences.

ELRA-L20 DST Dictionary (CORA)

Entries: 550,000 inflected forms.
Language: French
Format : ASCII
Medium:: tape, CD-ROM

Simple forms are divided into: 43,000 common nouns, 10,938 proper nouns, 19,500 adjectives, 8,150 noun-adjectives, 6,800 verbs, 6,200 compound nouns, 4,680 adverbs and adverbial phrases, 3,292 unelided words, 903 prefixes, 682 abbreviations and measures, 218 pronouns, 248 conjunctions and subordinating conjunction phrases, 186 prepositions and prepositional phrases, 86 determiners, 16 predeterminers, 14 co-ordinating conjunction phrases, as well as all possible homographs.

The DST includes semantic (cars, places, wines, etc.), syntactic (gender, number, tense, etc.), morphological (lemma), lexicological (homographs) and more specific syntactical information (prepositions followed by an infinitive form, intransitive verbs with « avoir » or « être », etc.).

ELRA-L14 Adverbial Equivalence Dictionary (CORA)

Entries: 1,200
Language: French
Format : Word Processing file (Word...)
Medium : floppy disk

Simplified equivalents for fixed expressions.

ELRA-T101 DESYN - Dictionnaire des SYNonymes (CORA)

Entries: Approx. 25,000
Language: French
Format: Tagged ASCII
Medium: Floppy disk

Card description : Thesaurus with substantives and adjectives, alphabetically ordered and cross-referenced (hyperonyms, synonyms or “see also”). Each entry consists of 3 or 4 cross-references.

ELRA-T103 Finance Terminology Database

This dictionary is a collection of sub-domains such as economy, commerce, business, banking, stock-exchange, negotiation, mail, conversation over the phone, etc. It consists of nearly 38,600 pairs of English terms and their French equivalents. It includes impersonal or conjugated expressions which turn out to be most useful in business. The dictionary also contains all kinds of locutions beginning with « at », « in », « on », etc.: « in blank », « on credit », « on trust », etc.

ELRA-L18 German lexicon (CORA)

Entries: 466,300. inflected forms
The same word can be represented in one or more files and thus counts for several entries.
Language: German
Format: ASCII
Medium: Floppy disk

This lexicon is divided into 7 main syntactic categories: nouns (97,000), verbs (236,200), adjectives and some adverbs (130,500), grammatical words (1,700), punctuation (40), prefixes (400), and suffixes (370). Each file consists of a word list corresponding to syntactical and morphological categories. The lexicon does not include lemmas.

ELRA-L15 Nominalisation dictionary (CORA)

Entries: 2,300
Language: French
Format : Word Processing file (Word...)
Medium : Floppy disk

Corresponding substantives for verbs.

ELRA MEMBERS AS OF 31/12/96

AFNOR, France - *Written*
Alcatel, Corporate Research Centre, Italy - *Spoken*
AP/HP - Service d'Informatique Medicale, France - *Written*
Aston University Birmingham, United Kingdom - *Terminology*
BJL Consult, Belgium - *Terminology*
Centre de Terminologie de Bruxelles, ILMH, Belgium - *Terminology*
Centre for Computational Linguistics, UMIST, United Kingdom - *Terminology*
CL Servicios Linguisticos, S.A., Spain - *Terminology*
CLIF (Research Community for Computational Linguistics in Flanders), Belgium - *Written*
Consorzio Pisa Ricerche, *Italy - Written*
Copenhagen Business School, Denmark - *Terminology*
Cray Systems, Luxembourg - *Spoken*
CSELT, S.p.A. (Centro Studi e Laboratori Telecomunicazioni), Italy - *Spoken*
CST (Center for Sprogteknologi) , Denmark - *Written*
Daimler - Benz, Germany - *Written & Spoken*
Defence Research Agency, United Kingdom - *Spoken*
Det Danske Sprog- og Litteraturselskab, Den Danske Ordbog, Denmark - *Written*
DFKI (Deutsches Forschungszentrum fur Kunstliche Intelligenz), Germany - *Written*
DIT (Deutsches Institut fur Terminologie), Germany - *Terminology*
Dragon Systems Ltd, United Kingdom - *Spoken*
Dublin City University, Ireland - *Written*
EDF, Electricite de France - Research & Development Division, France - *Terminology*
EP Electronic Publishing Partners GmbH, Germany - *Terminology*
EULINE, France - *Written*
Faculté Polytechnique de Mons, Belgium - *Spoken*
France Telecom/CNET, France - *Spoken*
Fry & Bonthrone Language Consultancy & Services, Germany - *Terminology*
Gesellschaft fur Terminologie & Wissenstransfer e.V. , Germany - *Terminology*
GSI-ERLI, France - *Written*
Handelshojskole Syd (Southern Denmark Business School) , Denmark - *Terminology*
IBM-European Language Business Unit, France - *Written*
IDIAP, Switzerland - *Spoken*
ILSP (Institute for Language and Speech Processing) , Greece - *Spoken*
INaLF, Institut de la Langue Francaise, France - *Written*
INESC, Portugal - *Spoken*
INFOTERM (International Information Centre for Terminology), Austria - *Terminology*
INIST (Institut de l'information scientifique et technique), France - *Terminology*
Institut fur Deutsche Sprache, Germany - *Written*
Institut National des Telecommunications - France Telecom, France - *Terminology*
Instituto Cervantes, Spain - *Spoken*
ITE (Linguistics Institute of Ireland) , Ireland - *Written*
IVNL (Institut voor Nederlandse Lexicologie), The Netherlands - *Written*
Jacobacci & Perani, S.p.A., Italy - *Written*
Johannes-Gutenberg Universitat, Institut fur allgemeine und vergleichende Sprachwissenschaft, Germany - *Written*
Katholieke Universiteit Leuven, ESAT - Spraak, Belgium - *Spoken*
Language Technology Group, University of Edinburgh, United Kingdom - *Written*
Lernout & Hauspie Speech Products, Belgium - *Spoken*
LIMSI (Laboratoire d'informatique pour la mecanique et les sciences de l'ingenieur), France - *Spoken*
Linguistique Communication Informatique, France - *Terminology*
MT News International, United Kingdom - *Written*
Philips Corporate Research, Germany - *Spoken*
Praetorius, Ltd., United Kingdom - *Terminology*
Rank Xerox Research Center, France - w*ritten*
Real Academia de Ciencias, Spain - *Terminology*
Real Academia de la Lengua Espanola, Spain - *Written*
Rigel Engineering, Srl., Belgium - *Spoken*
Siemens, AG, Germany - *Spoken*
SIETEC Systemtechnik GmbH & Co. OHG, Germany - *Written*
SPEX (Centre for Speech Processing Expertise), The Netherlands - *Spoken*
Sprakdata, Goteborg University, Sweden - *Written*
TELEFONICA I&D, Spain - *Spoken*
TERMCAT, Spain - *Terminology*
THAMUS, Consorzio per la Linguistica Computazionale, Italy - *Written*
Topterm VOF, The Netherlands - *Terminology*
UCL (University College London), United Kingdom - *Spoken*
Universitat Koblenz-Landau, Institut fur Computerlinguistik, Germany - *Written*
Université Catholique de Louvain, Groupe de Recherche Valibel, Belgium - *Spoken*
University of Amsterdam, Computer Centrum Letteren, The Netherlands - *Written*
VOCALIS Ltd., United Kingdom , *Written*
ZERES GmbH (Zentrum fur elektronische Ressourcen europaischer Sprachen), Germany - *Written*