The ELRA Newsletter

November 1998

Vol.3 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: Rébecca Jaffrain

Contributors:
Paul Baker
Lou Burnard
Khalid Choukri
Inge Gorm-Hansen
Edward Johnson
Bente Maegaard
Tony McEnery
Klaus Netter
Mark Sebba
Henrik Selsøe-Sørensen

ISSN: 1026-8200

ELRA/ELDA
CEO: Khalid Choukri
Assistant: Rébecca Jaffrain
55-57, rue Brillat Savarin
75013 Paris - France
Tel: (33) 1 43 13 33 33
Fax: (33) 1 43 13 33 30
E-mail: choukri@elda.fr
or jaffrain@elda.fr
WWW:
http://www.icp.grenet.fr/ELRA/home.html

Dear ELRA Members,

This is the last issue of the year, and we would like to take this opportunity to report on our own activities during the last fiscal year (October 97 - September 98), which will be discussed at our Annual General Assembly meeting to be held on 17 December 1998, in Paris. Of course, our members who are invited to attend the 1998 AGM will receive a more detailed report, including financial data, by mail.

This quarter has seen the end of our first project, LE1-1019, which was supported by the European Commission. The goal of the project was to build up ELRA's infrastructure. This project involved a commitment on our part to establishing an operational infrastructure able to address the technical, commercial, legal, and logistical issues related to the distribution of language resources. After three years of activity, ELRA and its distribution agency (ELDA) are showing their maturity. The low uptake of the first year is now far behind us: from distributing a mere 20 items in 1996, we improved our operations to reach over 180 in 1998.

This project (LE1-1019) enabled us to build the necessary infrastructure - its efficiency is one of our major concerns. We compiled a useful catalogue of marketable language resources (LRs) and drafted viable contractual schemes. These are used by ELRA in over 123 agreements with users and about 60 agreements with resource providers, not to mention the several dozen providers using such agreements for involvement in evaluation projects (such as *ROMANSEVAL*).

An important topic we planned to address was validation methodologies and validation procedures, and capitalising on on-going CEC-funded projects. Close cooperation with the *SpeechDat* project(s), and the SPEX Validation Center allowed us to produce a first draft of a validation manual for spoken language resources. Cooperation with EAGLES and PAROLE allowed us to draft first releases of validation manuals for both written lexica and written corpora. These manuals are all available via the ELRA Web site. Moreover, a first implementation of corresponding validation procedures is underway for the Danish, Italian, and Spanish lexica produced within the *PAROLE* project.

In order to be able to apply such validation procedures, ELRA will issue a call to set up a network of Validation Units within the next quarter. The call will be widely disseminated.

As you know, validation of resources and validation methodologies will have an important impact on the specification of new resources.

What about the future?

ELRA finances are more solid than they were a year ago, which permits us to face the future with more optimism and serenity on the one hand, but also with more ambition on the other, and there is still a long way to go. We need to work out agreed-upon and usable standards. We need to test the validation procedures (which have to be easy to use and efficient). Moreover, we need to have the right resources for R&D and commercial purposes, and they need to be well-adapted for system training as well as testing and evaluation. And we need to be more involved in on-going or upcoming evaluation programmes.

This issue begins with a paper on evaluation as discussed in Granada. Bente Maegaard elaborates on evaluation methodologies. Ed Johnson and his colleagues tell us about the LinguaNet prototype messaging system, a project that uses multilingual lexica resources to handle the needs of police staff across Europe. Klaus Netter presents two European projects, Pop-Eye and Olive, which deal with archiving film and video material for subsequent re-use. The system developed in these two projects can be regarded as the fully functional multilingual, multimedia information-retrieval system. The last paper comes from Tony McEnery and his colleagues, who describe the *MILLE* project, which highlights the importance of addressing the needs of minority languages.

This issue also includes brief descriptions of the latest resources we have secured for distribution, which are the following:

- speech resources: SIVA (Speaker Identification and Verification Archives), Chilean Spanish FDB-500, ILE: Italian Lexicon, MULTEXT Prosodic database, French Speechdat(II) FDB-1000.
- written resources: MULTEXT JOC Corpus, ARCADE/ROMANSEVAL corpus, MULTEXT Lexicons.
- terminological resources: Newbase, hydrogeology database, pedology database.

As this is the last issue of the year, we would like to take this opportunity to wish you a Merry Christmas and a Happy New Year in 1999.

Antonio Zampolli, President Khalid Choukri, CEO

ELRA Annual Report 1998

Khalid Choukri, ELRA/ELDA

The ELRA project objectives are derived from the *European Language Resources Association* (ELRA) mission. ELRA devotes its efforts to the fulfilment of language engineering requirements in terms of the availability of language resources (LRs). In order to do so, ELRA identifies needed resources, tries to enter into agreements with the right holders, and makes the LRs available to technology developers. ELRA's main tasks are therefore of a technical, legal, commercial, and logistic nature. ELRA also collects information about the market through surveys and feedback from its members. The Association organised the LREC conference (ELRA International Conference on Language Resources and Evaluation) and its satellite workshops which addressed evaluation issues and the problems of sharing LRs.

Summary of 1998 activities

The major achievements of 1998 cover several areas. We improved our distribution of language resources (LRs) by over 577% as compared with 1997; identified new LRs (securing over 100 new databases, mainly spoken and written); and disseminated information through the huge ELRA conference (LREC, over 500 attendees) and the regular ELRA quarterly newsletter (4 issues), as well as our membership drive. Our call for proposals to commission the production of LRs attracted many proposals and inquiries (about 30) of which 9 proposals fulfill our formal requirements and are under discussion. ELRA carried out the preparation of validation manuals to assess the quality of language resources and made these manuals widely available. We also started a pilot application to implement such validation for some of the resources produced by the PAROLE project. As it stands today, our financial resources will allow us to plan for important investments in the co-production of new language resources in order to make us self-sufficient for the next 5 years.

Distribution of language resources

Major efforts were devoted to the distribution of language resources which led to an 577% increase in our 1997-98 sales over 1996-97. Sales amounted to over ECU 700 thousand with 179 items sold in 1997-98, compared with ECU 120 thousand and 31 items sold in 1996-97. Despite our marketing and commercial efforts, we are still making most of our income from spoken language resources: 86.7% of our revenue was generated by speech products, 13% by written resources, and 0.3% by terminology resources compared with 87.3%, 11.9%, and 0.8% respectively in 1996-97. Most of our customers join ELRA before buying the LRs (which is justified by our pricing policy). Our contribution to the development of research activities has seen considerable growth as evidenced by the 1000% revenue increase highlighting the acquisition of expensive resources for research purposes. Our involvement in research and commercial developments is balanced and shows an increase of over 573% in terms of items distributed for R&D and 581% for commercial use. When it comes to the distribution by geographical area, we can see that this development is more significant outside Europe, both in terms of quantity of resources sold and in terms of revenues (increases of 900% and 733% respectively outside of Europe, compared to 445% and 460% within Europe).

Identification of language resources

In terms of our language resources identification task, we managed to enter into several new agreements and to increase our catalogue entries. The catalogue issued in September 1998 consisted of 105 speech databases, 17 written corpora, 47 monolingual lexica, over 125 multilingual lexica, and about 361 terminological databases, compared to 64, 15, 40, 69, and 361 respectively in October 1997. The agreements secured so far were with 21 providers of spoken language resources, 31 of written resources, and 8 for terminology databases compared to 19, 27, and 8 respectively for 1997. It is clear from this that we devoted more efforts to distribution, but it also indicates that we identified the key players in our fields very early in our operations.

Validation of language resources

Our involvement in validation and quality assessment has seen the release of validation manuals in the area of speech and written resources; these manuals were made widely available. ELRA started a pilot application to implement such validation for some of the resources produced by the PAROLE project (mainly lexica and written corpora for 3 different languages). The validation manuals will allow us to set up our network of validation units and other technical centres early in 1999. A call will be issued before the end of 1998.

Commissioning the production of language resources

Our call for proposals to produce language resources ("ELRA commissioning of new resources") attracted many proposals and inquiries (about 30). So far only 9 proposals fulfill our formal requirements and are under discussion. These involve multilingual terminology resources, SpeechDat-like databases and children's speech (3 to 7 year olds, in English, Spanish, and French), updates of existing dictionaries, parallel corpora, etc. We are still expecting further justification about the market/users for such resources.

ELRA Membership

Following the membership drive, we managed to attract several new members. We now have around 80 members (compared to 75 last year). What is noticeable is the increase of subscribers (members from outside Europe). This year 8 subscribers from Japan, the USA, and Canada joined the Association.

Promotion and awareness

Our contribution to information dissemination activities consisted of a very impressive international conference, the 1st ELRA International Conference on Language Resources and Evaluation – LREC. LREC and its satellite workshops, held in Granada from May 25 to June 1 attracted over 500 attendees. The programme committee selected about 197 papers. Eight pre-conference workshops and a major post-conference workshop about trans-atlantic cooperation (called Multilingual Information Management) were also organised. The conference included an industrial exhibition which enabled some of our partners to show their latest products.

One of the other means to make ELRA more visible consisted of our quarterly newsletter, issued in French and English. A special issue was devoted to LREC with summaries of the opening, closing and several technical sessions. We have recorded a larger number of visits to our Web site and the site has been updated on a very regular basis, both with new resource descriptions and documents of interest to the language engineering community, such as validation manuals. In order to promote ELRA, the Board members and the ELDA staff attended several conferences (both academic and business-oriented) and gave several talks highlighting our activities.

Future work

Today, our financial resources allow us to plan for important investments in the co-production of new language resources. ELRA will also consider new distribution channels such as electronic commerce and continue its process of validating some key resources (such as Parole lexica). We will also carry out quick and very specific surveys to identify the needs of our members and customers. Furthermore, ELRA will implement the joint-venture policy decided by the Board and set up an investor group consisting of the major players in our field which will help us to be more reactive in addressing identified needs. ELRA is and will continue to be involved in several "evaluation" projects and activities as a data supplier, or as the distributor of the data and know-how gathered in the course of such activities. ELRA will also continue to contribute to the debate within ELSE (proposal for a European infrastructure for evaluation) and to supply data and other relevant information to projects dealing with evaluation such as the following: AURORA (developing draft standards for distributed speech recognition (DSR) which will be standardised by ETSI), Romanseval/ Senseval (multilingual text alignments), Amaryllis-2, etc. In most cases ELRA will provide the raw data and will distribute the processed data to the participants for the evaluation process.

We will of course continue our regular activities such as identifying new resources, issuing the four editions of the newsletter (in French and English), and so on. We will focus more on the services offered to our members through the Web, focusing in particular on an ongoing project to improve our catalogue which will see results within the next quarter.

Early in 1999, the ELRA Board and the LREC programme committee will start their discussions about the LREC 2000, planned for May or June 2000.

Further information

More information about ELRA is available on our Web site, including our catalogue of language resources, generic contracts to be used when brokering language resources, validation manuals for spoken and written databases, etc. The ELRA newsletter, including previous issues, is available on request. The proceedings of the LREC and the satellite workshops are available at ELRA's offices.

Evaluation methodologies

Bente Maegaard, Center for Sprogteknologi

The LREC (*Language Resources and Evaluation conference*) in Granada in May 1998 showed an enormous interest in evaluation of language technology, speech technology etc. as a discipline in itself. Similarly, evaluation methodological issues are coming up in almost any conference concerned with language and speech technology, not just as an aspect of a project or a proof of an approach ('this is what we did, and this is how well we performed'), but also as methodological considerations cutting across specific project developments. Below, we briefly describe the EAGLES NLP Evaluation approach and a few projects which have used EAGLES or similar approaches.

The *EAGLES* project, started 1993, focuses on providing standards for various aspects of language technology, one of these being evaluation of NLP. Since the objective is to develop standards which will be widely accepted, ISO was a good point of departure. Consequently, the EAGLES evaluation group has been using the ISO 9000 series as inspiration and has been further developing in particular ISO 9126. (Though a side effect, it has been a pleasure to see that some of the ideas developed in EAGLES have also been developed in ISO and will become part of the revision of ISO 9126 which is underway). The methodology developed can be used for products, *adequacy evaluation*, and for projects, *progress evaluation*. We have been focusing on adequacy evaluation.

There are several aspects of the EAGLES methodology which we find important. Below we shall briefly mention two: the user-centeredness and the formalisation aspect. For more information, see *http://www.cst.ku.dk/projects/eagles2* where you also find a discussion forum on evaluation.

User-centred

The EAGLES evaluation group decided to make the user requirements and ways of expressing them a central theme. In fact, we need evaluation only because of the users. But already here, it is realised that there are many different users of an evaluation, so we have to be quite precise when we talk about users. Users of an evaluation may be developers, providers, funders (these are all at the 'production end'), as well as managers, end-users, consumer magazine employees (being at the 'consumer end'). In each case the user requirements have to be specified. The EAGLES methodology is broad enough to satisfy all types of users, but in the work we have been focusing on the end-user since we were also focusing on existing products on the market. Furthermore, even within the class of end-users, different users with different tasks require different performances from the system, so in each case a detailed formalised description of the user requirements has to be made.

Formalisation

Probably the most important extension of the ISO standard is the formalisation of descriptions of products and of classes of users. These descriptions are expressed in terms of feature structures. There are two reasons for striving for formalisation. First of all, formalisation facilitates standardisation. Secondly, formalisation facilitates automation, and automation makes the testing phase easier and more reliable. So, even if formalisation and automation are not always possible, it is an ideal goal to strive for, and experience shows that with ingenuity one can get quite far.

Other projects

The two projects we mention below are not principally concerned with evaluation, but evaluation has become an important task in both cases.

The *MULTIDOC* project is an EU project in the field of multilingual automotive product documentation. This project uses an evaluation methodology which, like the EAGLES methodology, is highly inspired by software evaluation and assessment and hence by the ISO 9000 series. The MULTIDOC approach focuses on diagnostic evaluation, i.e. evaluation to be performed throughout the project development, to ensure that development follows user requirements, and to detect errors and deviations. The project separates the evaluation of software systems and lingware where lingware again is broken down into resources such as grammars and lexica, and technologies such as analysers, translators and generators.

Evaluation has been taken very seriously in the MULTIDOC project which provides a very good example of the advantages of rigorously taking the user requirements as the most important point of departure for software development.

The EU project *ARISE* (Automatic Railway Information Systems for Europe) has used the EAGLES methodology for their user validation. ARISE aims at providing callers with information about train schedule by telephone. The project covers the Dutch, French and Italian languages. The system is aimed at handling the bulk of routine enquiries automatically - there are 200 million calls annually to European railway centres, of which 20% currently go unanswered due to the cost of manual service. The interested party - the user - is the railway company in this case. The railway company wants to provide this service as this is a way to sell more tickets. Of course the caller is a user as well, and the caller satisfaction influences the user satisfaction, but it is important to keep in mind who the main user is, and the ARISE project is quite clear on this.

The ARISE validation viewpoint is interesting as we have normally been working with the end-users' viewpoint in EAGLES (and this is also the case for MULTIDOC). The work by the ARISE project shows that the EAGLES methodology is applicable also in this case. Furthermore, the formalisation requirements of the EAGLES approach has actually helped the ARISE project to think about the user validation. Basically: What are the user requirements? And in what way do the systems respond to those requirements? The user requirements are broken down into four main objectives. The railways want 1) to provide a service (information), 2) to have this service accepted, 3) at a reasonable price, and 4) for the right type of callers. The ARISE project examines to what extent the systems built respond to each of these requirements. The EAGLES project has cooperated with the ARISE project by providing feedback to an earlier version of the validation document, and this fruitful cooperation has led to the organisation of a workshop in April 1999 (see below).

Evaluation and validation

In the MT community there is a growing consensus that some classification of products is needed in order to guide end users. This need is arising as language technology products are reaching the mass market. The discussion started at the *MT Summit* in San Diego 1997 and was continued at the AMTA conference *Machine Translation and the Information Soup* in Langhorne, Pennsylvania, in October 1998. As mentioned, the aim is to give a classification so that the user can distinguish an MT system from an electronic dictionary, but there is no

doubt that in order for such a classification to be of any interest, it has to be combined with some level of evaluation.

Validation discussions have been going on in ELRA as well, and now validation manuals are available for lexica, corpora and speech. Validation of a language resource basically consists in checking that this resource is what it claims to be: a dictionary should exhibit the most important features of a dictionary (e.g. words and part-of-speech), and it should conform to its own specifications (e.g. all words have part-of-speech). As can be seen, this type of validation is very close to the classification checking for MT mentioned above, and we will certainly see more of this 'basic' or 'low-level' validation for the consumer market, presumably gradually enhanced with real evaluation.

Workshop on evaluation

A two-day workshop *European Evaluation of Language Systems* (EELS), bringing the EAGLES evaluation approach to practical work, is being organised by the company Compuleer and the EAGLES Evaluation group. The workshop will give a practical guide to using the EAGLES methodology. Apart from teaching the general methodology, it features case studies and special tutorials for software developers and linguists working in language technology industry. The workshop takes place in Hoevelaken, The Netherlands, 12-13 April 1999. Further information: Marc Blasband, cplr@worldonline.nl.

References

EAGLES Evaluation of Natural Language Processing Systems, Center for Sprogteknologi, Copenhagen, 1996. Also available at *http://isscowww.unige.ch/projects/ewg96/ewg96.html*.

Lise van Haaren, Marc Blasband, Marinel Gerritsen & Marcha van Schijndel: *Evaluation Quality of Spoken Dialogue Systems: Comparing a Technology-focused and a User-focused Approach*. In: LREC Proceedings, Granada, 1998, p. 655-659.

Margaret King., Bente Maegaard: *Issues in Natural Language Systems Evaluation*. In: LREC Proceedings, Granada, 1998, p. 225-230.

Jörg Schütz, Rita Nübel: *Evaluating Language Technologies: The MULTIDOC Approach to Taming the Knowledge Soup*. In: AMTA'98 Proceedings, Machine Translation and the Information Soup, Springer 1998, p.236-249.

Bente Maegaard
Center for Sprogteknologi (CST)
Njalsgade 80
DK-2300 Copenhagen S - Denmark
email: bente@cst.ku.nl

POP-EYE and OLIVE - Human Language as the Medium for Cross-lingual Multimedia Information Retrieval

Klaus Netter[1], DFKI

Introduction

Archiving quite obviously plays a central role for the reuse of film and video material. In this process, the detailed and comprehensive documentation and profiling of the archived material is a prerequisite for an efficient and precise access to the data. While in the domain of textual digital libraries, advanced methods of information retrieval can support such processes, there are so far no effective methods for automatically profiling, indexing, and retrieving image and video material on the basis of a direct analysis of its visual content. Although there have been some advances in the automatic recognition of images, these are still so limited that they will not provide a sufficiently robust basis for effectively profiling large amounts of visual data.

In this paper we present two European projects, *Pop-Eye and Olive*[2] which attempt to address this problem. The projects are centred around the assumption that, due to the non-discrete nature of visual data and content, for a considerable amount of time the automatic indexing and retrieval of image and video material will only be possible on the basis of human language as the medium for profiling and for searching. Accordingly, the profiling processes employed in these projects, which are required for constructing detailed indices, take into account all different kinds of linguistic material associated with a video production, such as subtitles (close and open captions), written transcripts, the spoken word, or background material, such as production scripts or press releases. On the basis of normalised textual data, indices are built which allow to access productions not just as a whole but relative to shots or short sequences directly related to index terms. Through the use of automatic translation technology, the user can search and retrieve material in different languages, such that full cross-lingual access is provided.

In the following, we first briefly describe the projects from the users' point of view, i.e., we sketch some typical archiving and documentation processes, and the requirements and needs to be met by a digitised video archive. We also give an indication of the kind of linguistic data that are typically associated with video productions and which can be used for the indexing and profiling of the material. We then present an overview of the functionality of the system(s) developed in the framework of the projects Pop-Eye and Olive, showing how the cross-lingual access to multimedia data is realised, and finally give a brief project and implementation note.

Archiving and reuse of video productions

The primary users of the Pop-Eye and Olive projects are major European Television Stations, comprising ARTE (Strasbourg, France), BRTN (Brussels, Belgium), SWR (formerly SWF, Baden-Baden, Germany), and TROS (Hilversum, The Netherlands), as well as the French national audio-video archive, INA/Inathèque in Paris, France, and a large service provider for broadcasting and TV productions, viz., NOB in Hilversum, Netherlands. For all of these institutions archiving of video productions plays an important role, be it for the purpose of re-broadcasting or reselling existing productions, for reusing part of the material in new productions or for generally supporting research in video data bases. In particular, the latter two functions make it very important that the archives' customers have maximally detailed access to the content of the video material. Reusing parts of existing material can reduce the production cost considerably and therefore makes it highly desirable that the full and detailed content of a video be documented and accessible without having to view the entire video.

However, developing the necessary content descriptions for video productions manually (or rather intellectually) is an extremely costly and labour intensive enterprise. On the average, one has to assume that a trained documentalist can describe video productions at a ratio of 1:10 and higher, i.e., for one hour of video material at least ten hours of human labour have to be calculated. Even for larger institutions, this makes it almost impossible to provide the necessary profiling for all productions. As a consequence, the archiving is often limited to capturing the factual data together with some few keywords. Some notable exceptions are among others the quite sophisticated documentation provided through the FESAD database in the German ARD federation or the content-related disclosure of the BRTN video archive.

For production and research purposes, ideally one would be allowed to access the digitised video material online through some Intranet or even the Internet. For example, a producer should be able to log into the digital video library, submit a request, browse through the data descriptions and then download and view the relevant sequences. This would mean that he could put together in his workplace a cut list with the sequences and programmes which he wants to obtain physically from the archive.

Automatic Indexing and retrieval

To answer such problems and demands as just described, Pop-Eye and Olive attempt to provide on-line access to video material on the basis of linguistic material associated with the data. The tasks performed by the system(s) are the following:

- Video material and linguistic data are digitally captured and aligned with each other, where necessary, by inserting the time code of the video into the textual representation.

- The texts are processed on the basis of state of the art language technology and different indices are constructed from the text. Where possible the texts or the indices are translated.
- In response to a search term, the system provides the user with pieces of texts matching his query, and allows for downloading and viewing the corresponding video sequence via the time code.

Among the capturing steps, the digitisation of the video material is necessary since it is currently still only very rarely available in a digital format. By reducing the size, resolution, number of colours or frames the digitised version can be sufficiently compressed to allow for an efficient downloading across some network. An interesting alternative to such reductions is the automatic derivation of an image-based story board, as it is developed, for example, by the Euromedia project (http://www.foyer.de/euromedia/). Through such a story board, a video sequence is represented as a succession of still images, each representing a different shot or at least different angles in a shot.

The linguistic data associated with a video basically come in two classes. They are either time coded directly or inherently contain some time code, or they are textual representations without any time code. Among the former are above all subtitles, i.e. close captions which are typically abbreviated text representations in the same language as the spoken word, serving mainly for better understanding for the hard of hearing, or open captions which are translated subtitles of the original. Since subtitles are typically time-coded text files, they can be processed and indexed like any other text file, providing keys into the content of the video. The second type of linguistic data inherently linked to the temporal sequence is of course the spoken word itself. To capture this stream, automatic speech recognition and transcription can be applied. Currently, speech technology is still somewhat limited and does not guarantee completely reliable domain- and speaker-independent recognition. However, it has to be kept in mind that for the purpose of indexing and retrieval, a 100% recognition rate is not absolutely necessary, since not every word will have to make it into the index, and not every expression in the index is likely to be queried. In addition, speech recognition can also be used as a secondary means to support automatic time coding of the second class of data, as for example manual transcriptions, which have to be carried out as a first step in the subtitling process or for the purpose of translation. If such data are available,[3] the cleaner and more reliable transcriptions can be used as the basis for indexing. The necessary time-coding can then be derived by automatically aligning the result of speech recognition with such a transcription. Basically the same method can be used if there are production scripts or other types of descriptions reflecting the time line and the spoken word.

The second major step in the (off-line) processing is the analysis of the written texts and the construction of indices. While most practical approaches to information retrieval build on very little linguistic knowledge and mostly rely on purely statistical methods, the projects attempt to combine the linguistic and statistical approaches. This means, for example, that every text can be analysed with shallow linguistic processing technologies, which account for rule-based lemmatisation, part-of-speech disambiguation, or the mark-up of phrases, which can then be extracted as index terms. There are also analysis methods being developed which go even further and allow, for example, to identify proper names, to determine the head-modifier structure of term expressions, to extract specific terminology, or to establish conceptual relations. It is quite essential, however, that these techniques are supported by powerful statistical methods, such as vector space modelling, which provides a measure for the similarity between two or more pieces of text, or by fuzzy indexing, which guarantees the necessary robustness by abstracting from the surface form, and which makes sure that a query term can be matched with all kinds of variants found in the index.

Different technologies are employed to provide multi-lingual access to the textual data, all of which build on off-line translation rather than on-line translation. In Pop-Eye and Olive, following an approach developed in the project Twenty-One (http://twenty-one.tpd.tno.nl/), the original texts are either fully translated (by means of the Logos translation server), or they are partially translated at least as far as the index terms are concerned. This means that the monolingual indices which can be constructed (and searched) cover the full multilingual document base, thus allowing reference to documents in different languages.

There are of course several other systems which make a claim to providing multi-lingual information retrieval, such as the combination of Altavista and Systran, or the Coronado system by L&H, which is built according to the same model as developed by the Mulinex project (http://www.dfki.de/lt/projects/mulinex/). However, in particular, the Altavista/Systran combination suffers from the problem that the user has to know the foreign language in order to formulate his search and find a relevant document. Only after the retrieval of the foreign language document can he ask this document to be translated into his own language. Systems like Mulinex and Coronado, on the other hand, help the user to translate his query into other languages, but then retrieve only the original in the foreign language.

In the final on-line querying and retrieval step, the user then enters his query in his own language. Normally there is no need for the translation of the queries, as the indices have already been translated into his language. As a response, the user first receives some pieces of text which match his query together with the relevant identifiers specifying the corresponding video or the time code referring to a video sequence. The texts can be, for example, phrases or full subtitles, and they can be originals or automatically translated. The user then has the option to view more textual information, e.g., the subtitle sequence in the context, the text in its original language, or he can directly download and view the relative video sequence, or a story board representing this sequence.

In its full extension, the system developed by Pop-Eye and Olive can thus be seen as one of the first, if not the only fully functional multilingual multimedia information retrieval system, which covers and handles all possible kinds of different media, ranging from speech, via text to images and video. However, it should be clear, of course, that the discourse and linguistic data associated with a video will not always be a direct reflection of the images and the visual content of the video. In particular, there will be a broad range of variation between more descriptive texts, like documentaries, where the commentary refers to and explains the visual content, and programmes of the drama type, where the dialogue and discourse complements the visual content. Thus, the approach taken in the two projects will have some clear limitations, and future experience and evaluation will have to show for what type of programmes the approach is most suitable.

The system is implemented through the co-operation of several technology providers, research institutions and universities. These include TNO-TPD Delft, which built the core indexing and retrieval functionality, VDA BV Hilversum, which is developing commercial software for the TV sector and which built the video capturing software and is responsible for system integration, the University of Twente and the LT Lab of DFKI GmbH Saarbrücken, which are responsible among others for the language technology, the University of Tübingen, carrying out the evaluation in Pop-Eye, CNRS LIMSI and Vecsys SA Paris which are developing and integrating the speech recognition modules, respectively.

Notes:

1. I gratefully acknowledge the contribution of Joop van Gent and Wessel Kraaij (TNO-TPD), Franciska de Jong (TNO/University of Twente), Godfrey Smart and Wim van Bruxvoort (VDA) and Jean-Luc Gauvain (LIMSI), to mention just a few who crucially influenced and shaped the design of the systems developed in Pop-Eye and Olive.

2. Pop-Eye (LE1-4234) and Olive (LE4-8364) are both funded by the European Commission under the Telematics Application Programme in the Language Engineering Sector. Pop-Eye started in 1997 and will last until 1998, Olive in 1998 lasting until 2000. The overall budget of the two projects together is 3.8 MECU. The languages covered by the two projects are Dutch, English, French and German.

3. In the case of a bi-lingual stations such as ARTE, this is a prerequisite for almost all programs which are broadcast in the two languages German and French.

Klaus Netter
Language Technology Lab
German Research Center for Artificial Intelligence DFKI GmbH
Stuhlsatzenhausweg 3, D-66123 Saarbrücken, Germany
E-mail: Klaus.Netter@dfki.de
URL: http://www.dfki.de/~netter/

LinguaNet? We Need it Now: Delivering Multilingual Messaging and Language Resources to the Police

Inge Gorm-Hansen, Edward Johnson, Henrik Selsøe-Sørensen, Copenhagen Business School (CBS)

Key Words: operational languages, multilingual messaging, police, emergency services, terminology, mission critical, normalisation.

This article briefly introduces the LinguaNet prototype messaging system recently tested by units of the European police community. Engineered for mono- and multi-lingual, mission critical communication, its objective is to allow reliable conversion between languages during real time technical co-ordination activities. The basic principles apply wherever a high proportion of the core textual, graphical or acoustic content is predictable and controllable as in business dealings, medical communications and distributed manufacturing.

A seven-language template version of LinguaNet is now installed at 36 operational police sites in seven European countries and is in use on a daily basis.

The (societal) problem tackled

European police officers investigating credit card fraud, vehicle theft, missing persons or involved in a cross-border incident in progress must be equipped to make direct contact internationally. The lifting of internal border controls across Europe has increased the need to find solutions to this problem as there is evidence that criminals are increasingly exploiting weaknesses in police communications to commit crimes both within the Community and across its external border.

User requirement

- safe,
- reliable,
- point-to-point,
- easy to install,
- easy to use,
- portable,
- low training costs,
- inexpensive to purchase and run, able to use available connection e.g. PSTN in first instance,
- user specified templates for operational messages, translation modules for templated/ controlled text,
- interfaces in all user languages,
- able to be upgraded (functionality),
- able to be expanded (languages and sites),
- able to carry graphics,
- able to carry sound files,
- assemble attendant multilingual police lexicons,
- assemble attendant police specific databases, direct connectivity to national criminal databases,
- transferable to multi-agency multi-national disaster scenarios.

Origin

The present system and network grew from a small UNIX-based, PolyML prototype created in 1994 by Prolingua Ltd. for a group of French, Belgian and British police units. That prototype itself, together with an analysis of police communications for the combined Anglo/French policing of the Channel Tunnel, produced the multilingual police messaging corpora from which the message types (wanted or missing persons, vehicles, accidents, bankers cards, firearms etc.), lexicons and protocols for the present version were derived.

In most cases normative procedures imposed upon the message structure, data elements and lexicon were sufficient to effect consistency and language conversion without compromising the communicative potential of the messages. Indeed normalisation engendered confidence and thus improved communicative value[1]. The addition of pictures and other graphics further enhances the messages and acts as a reciprocal gloss to the text elements.

Lexical "discrepancies"

Even in this highly disciplined and seemingly highly predictable domain, the creation of linguistic parities for simple messaging is fraught with difficulty, as is the production of useable multilingual lexical resources for reference purposes.

"Køretøj" translates as "vehicle", "véhicule", etc., but "rigspolitichefen" has no equivalent in the British nor in the French police organigrammes. Even so, users need to get from the specific national phenomenon to the closest foreign equivalent and tend to consider this only as a matter of translation. The definition of "closest" is often context-dependent. The planning of a lexicon/knowledge base for operational use must take the huge number of such cases into account.

A Danish police officer who looks up what "køretøj" is in French expects to refer to an object also known to the French colleagues. In the case of "rigspolitichefen", the average lexicon user will have the same expectation - not realising the basic pragmatic reasons for the inevitable discrepancy.

Three typical situations may cause a user to look up "rigspolitichefen" in English, for example:

Case 1: A Dane wants to get in touch with his British counterpart.

Case 2: A Dane wants to explain to an English speaker who the latter is dealing with.

Case 3: A non-Danish speaker may look up "rigspolitichefen" in order to find out who he is dealing with or reading about.

In case 1, a typical operational context might be where a Dane requests contact with the British counterpart of "rigspolitichefen" because "rigspolitichefen" would be the right person to contact in that matter at the Danish national level. The objective, however, in this case is not to identify the person, but to find out exactly to which authority the request has to be addressed.

In case 2, the Dane - as an active user - expects the translation to transfer sufficient knowledge about the Danish administrative structure to enable the British colleague to place "rigspolitichefen" and deduce from that how to interact with him. A translation, e.g. "Danish National Commissioner", will call upon the reader's knowledge about commissioners and whatever that entails, but given that the UK police structure does not have (except in the case of the London Metropolitan Police) a commissioner at all, the Brit may not be able to interact appropriately after all.

Case 3 is parallel to case 2 in regard to the objective of the translation, the difference being that case 3 is seen from the perspective of the passive user.

It is obvious that the issue of the active versus passive use of bilingual dictionaries is raised at this point. In this case, however, it becomes even more complicated because of the multilingual vocation of LinguaNet.

In an operational context, it is clearly case 1 which is of major importance. We proceed on this basis.

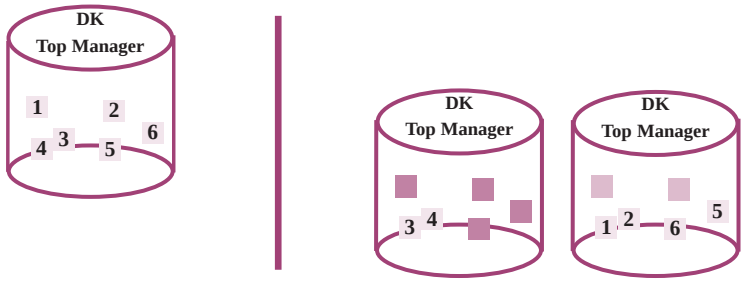
Let the barrel to the left on the picture next page symbolise "rigspolitichefen" who is in charge of matters 1-6; let the vertical black line be the language barrier and let the two barrels to the right be the British structure, which is less centralised.

The Danish user sets out to get a translation of what he knows as the top manager, i.e. "rigspolitichefen" because he thinks a translation into English will teach him whom to deal with in a given matter, e.g. "matter 5". The Danish user should now ideally learn from the lexicon:

a) that on the British side of the language barrier, there is no equivalent counterpart because the British system is different from the Danish one.

b) that instead of a translation he needs to find the appropriate person or unit in charge of "matter 5".

Cases like this are difficult to handle in an efficient and friendly way, especially when more than two countries/languages are involved. In such cases normalisation procedures do not work well and messaging must be supported by various knowledge base types.



Database reference lexicons

Useful database facilities such as multilingual police lexicons[2] and police facts sources were developed as "add-ons" in tandem with the messaging system built[3] by researchers at CBS using the data mining tool INTEX text analyser and System Quirk. MT tools for free text segments of messages were also tested by these researchers.

The team engaged on this project were given access to existing (de-sensitised) police corpora. The overall aim was to create and demonstrate multilingual police resources which could be made available to this and possibly other systems. Topics covered in this manner included:

- person description database (Danish, English and French Multi-term)
- justice and home affairs database (9 EU languages Multi-term)
- drugs database (English, French, German, Spanish and Danish Multi-term)
- police facts database (Danish, English, French and German sources Multi-term)
- multimedia firearms database prepared by Swedish police.

Technical details

The current LinguaNet software comprises a front-end application program providing message manipulation and translation together with an optional back-end to enable peer-to-peer message transfer. It runs in Windows™ 95 or Windows™ NT 4.0 and is compatible with Windows™ Messaging and Microsoft™ Mail. It has the standard Windows™ GUI with icon toolbox of commonly used forms. The messages may contain graphical information from scanners or digital cameras. JPEG picture compression reduces storage space and transmission time. The LinguaNet Service Provider for MAPI provides peer-to-peer connectivity over PSTN, ISDN or GSM.

A recurrent enquiry made by operational police users concerns speech technologies. This has stemmed from a style of work requiring maximum mobility and minimum electronic clutter. It is clear that the use of domain specific normalised vocabularies suit available speech technologies for "hands free" form filling; speech output to mobile radios and speech elicitation of selected message segments (output in speech).

Additional languages and applications

The addition of further languages to LinguaNet templates does not require the large developmental overhead normally associated with "whole" grammar translation systems and can be applied rapidly therefore (within a worked domain) to a large number and variety of national and minority languages. More extensive grammatical and other solutions to the communicative barriers will be implemented on the back of this already functional installation.

Much of the LinguaNet work benefits from the prior experience of the team in the design of operational languages. These are subsets of natural language deployed wherever a language circuit is a requirement in the co-ordination and control mechanisms of a technical operation[4]. Examples are the co-ordination of ships, aircraft, trains, spacecraft, police, fire brigade, and ambulance services communication.

In response to the recognition that full natural language is an unreliable tool, subsets have either evolved or been created. The Air traffic control language is an obvious example; sea traffic control and the protocols established in preparation for police and emergency service operations at the Channel Tunnel are others[5].

These languages typically address such problems as functional ambiguity, terminological imprecision, polysemy, inconsistency of alphanumeric data; random abbreviation, elision and ellipsis. Operational languages contain measures for countering such difficulties.

An example drawn from maritime radio communication illustrates some of the counter-measures taken. An utterance which in natural language might be any variant of "I'd like you to take the pilot from the SB buoy at 2 o'clock" must be (in Seaspeak) "Advice: meet the pilot, position: SB buoy, time: one-four-zero-zero UTC."

Results

The Consortium has had positive feedback in the users' own functional terms: criminals apprehended and property recovered. Reports from the LinguaNet frontier units include: the recovery of stolen vehicles to the value of ECU500,000 via just one LinguaNet terminal; a thwarted international child abduction in the Netherlands; stolen hire cars (already crated) intercepted at Brussels airport and the recent successful use of the system during one of the World Cup venues (Lens) to counter hooliganism.

Work has already started for public services involved in responding to major incidents such as maritime disasters, floods, chemical and nuclear accidents, major fires and aircraft crashes. In the United Kingdom it is the police force which organises a "Casualty Bureau" which compiles casualty details . It is expected that the instances where the agencies of more than one nation are involved in a disaster response and where victims come from several nations will increase.

Conclusion

The main achievement of Test-Bed LinguaNet has been to provide a simple but effective solution to an urgent need. Technical compatibility and openness has been sought so that the system may be enhanced, expanded, modified or embedded as operational requirements, new developments, alternative application sectors and the marketplace dictate. The system now in place and in use provides a powerful motivation and a convenient habitat for development.

Notes:

1. A principle put to good effect long ago in the EDI standards which lead to EDIFACT.

2. An English French Police and Emergency Services lexicon built for a previous project (INTACOM 1994) was integrated with the Unix version of the LinguaNet software.

3. A detailed treatment of the lexical and knowledge base issues is forthcoming 1998: *Keystone Ontologies for Cops.....*by the same CBS authors.

4. See Johnson E. Proceedings of the International Language Symposium Volume 4 Paris 1989 Les langues et la concurrence économique also: *Fachsprache International Journal of LSP Vienna 1-2 1990* Language and Economic Competition.

5. Seaspeak Reference Manual 1984 Pergammon Press E. Johnson, Lt. Alan Glover, Peter Strevens and Capt. Fred Weeks.

AirSpeak: Radiotelephony Communication for Pilots Prentice Hall 1988 F.A. Robertson and E. Johnson.

Police Communications and Language and the Channel Tunnel Policespeak Publications 1993 E. Johnson, M. Garner, S. Hick & D. Matthews.

Please address initial enquiries relating to this project to the Project Coordinator: Edward Johnson - tel:+44 (0)1223 276815; fax:+44 (0)1223 276813; email: ed@prolingua.co.uk

Minority Language Engineering

Paul Baker, Tony McEnery & Mark Sebba (Lancaster University), Lou Burnard (Oxford University Computing Services)

The Minority Language Engineering Project *(MILLE)* is a joint project between the Department of Linguistics at Lancaster University and Oxford University Computer Services, funded by the Engineering and Physical Sciences Research Council in the UK. It seeks to investigate the development of corpus resources for UK non-indigenous minority languages (NIMLs).

Obvious benefits of the creation of multilingual corpora to translators, lexicographers and dictionary builders are in the construction of bilingual dictionaries and aids for translators (e.g. construction of "technical" terminology for phrases such as housing benefit or visual display unit), leading to long-term improvements in the output of translation. To date, English, French, Spanish have benefited from the creation of bilingual and trilingual corpora such as CRATER (McEnery, Wilson, Sanchez-Leon & Nieto-Serrano, 1997). Similarly, many European languages have benefited from monolingual corpus construction. Our aim is to bring the benefits of mono and multilingual corpus construction to as wide a range of non-indigenous European languages as possible, starting with the UK.

In 1991 ethnic minorities accounted for approximately 6% of the population of the UK. Although the majority of residents in the UK speak English as a first language, there are large areas of the country where ethnic minorities cluster, forming considerable communities of speakers of such languages as Somali, Vietnamese, Cantonese and several languages from the Indian sub-continent (Gujarati, Punjabi, Urdu, Hindi, Bangla, Sylheti). Computationally, however, these language are ill-served. As noted by Somers (1997), beyond word processing and accompanying fonts, many of these "exotic" languages do not have adequate computational resources (e.g. spell-checker, style-checker, mono and bilingual dictionaries, thesauri, technical terminology management, CAT and MT).

We view it as a problem that resources and materials have not been produced to address translation tasks faced daily in urban Britain - translation into such languages as Hindi, Punjabi, Somali, Cantonese and Urdu. Our project aims to investigate the feasibility of creating NIML corpora by determining:

- the extent and availability of existing NIML data
- the requirements of language engineers who will need tools in order to exploit our corpora
- methods of putting the data into a machine readable, accessible format.

At the start of the project we also stated our intention to build two small parallel NIML-English corpora, as well as to investigate the feasibility of creating corpus resources for right to left languages (Urdu, Arabic) and for spoken NIML data.

Deciding which languages to focus on in creating corpora, and which data to use was initially problematic as we do not have the time or resources to construct corpora in every UK NIML. We decided to concentrate on one Indian language and one Chinese language, which were both used widely in the UK. In collecting corpus data we also limited our search to data which is either produced in the UK or produced for a mainly UK NIML audience. It would have been easy to construct corpora using e.g. overseas foreign language materials replicated on the World Wide Web, but as the target audience in most cases was not the UK, we decided not to take this approach.

At the time of writing (four months into the project) we have created small parallel corpora for Punjabi-English (modern children's stories) and Cantonese-English (Department of Health help leaflets), both of which have been encoded using a subset of the Text Encoding Initiative (TEI) Guidelines known as TEI-lite. Work on the other aspects of the project is on-going; we are in the process of contacting a number of UK local councils, translation and interpretation units, religious community groups and producers of foreign language media in order to determine the availability of electronically-occurring NIML language resources.

As well as applying TEI to NIML corpora, we are also investigating issues concerning the storage and exchange of electronic data in non-English scripts. Previously, work on languages such as French, Spanish and German have not found storage and exchange overly problematic - 8 bit character sets composed of 256 characters can handle accented Roman characters, while for 7 bit interchange, SGML entities such as & eacute; can be used to encode such characters. For Indian and Chinese languages the problem exists on a much larger scale. Many fonts exist for the representation of NIMLs, but not all corpus users may have access to such fonts. Also, in collecting corpora, it is likely that multiple sources will be employed, which may not all use the same font to encode their data. Unlike romanised fonts (e.g. Arial, Times New Roman) where an "a" key press (or ASCII code 97) will always give something resembling the small-case letter "a", with Indian language fonts pressing "a" on a Roman keyboard may result in a different Indian character appearing, depending on which font is being used. The Indian fonts map the Roman keyboard to Indic scripts in different ways. Finding a way to standardise this information so that the end user does not have to have access to multiple sets of fonts is one goal of MILLE.

So far, we have begun to examine two possible strategies; the first involves the creation of writing system declarations (or wsds) for each font used. Writing system declarations are TEI-conformant structures which document character representation, interchange and transliteration schemes. However, as they are used for documentation purposes only, they require some extra work by the end user - e.g. a program which will take the mapping information in the wsd and implement it accordingly. Another alternative is to make use of Unicode, a 16-bit character set, as the base character set for the corpus. This would allow interchange between languages using one character set only. However, at present there is a dearth of editors which are able to exploit Unicode to its full capability. Currently, we are examining the possibility of converting font-based representations of NIML scripts into Unicode using UniEdit (a Unicode-compliant editor developed at Duke University). Again, this work is still somewhat experimental - at present UniEdit's Indic characters are still in development. However, we are working closely with the Duke team and should be testing out this type of transfer before the end of 1998.

Finally, we are aware that the languages we are dealing with in the UK are also important in other European countries. We welcome feedback from other Language Engineers, and have constructed a Web site at http://www.ling.lancs.ac.uk/monkey/ihe/mille/public/title.htm. This web site contains a questionnaire you can fill in to tell us about what you see NIML language engineering priorities to be and what the NIML situation is in your own country.

Our project is still in its infancy, yet we feel confident that we have made a good start – however, we are breaking new ground in corpus construction and we are aware that there is still much to do!

Reference

McEnery, A.M., Wilson, A, Sanchez-Leon, F. & Nieto-Serrano, A. Multilingual Resources for European Languages: Contributions of the CRATER Project, Literary and Linguistic Computing, 12:4, 1997.

Somers, H. Machine Translation and Minority Languages. Translating and the Computer 19. Papers from the ASLIB Conference 13/14 November 1997.

For more information, please contact:
Tony McEnery
Lancaster University
Email: mcenery@comp.lancs.ac.uk

New resources

Keys: R: for research use - C: for commercial use

ELRA-S0028 SIVA (Speaker Identification and Verification Archives)

The Italian speech database SIVA (Speaker Identification and Verification Archives), is a database comprising more than two thousands calls, collected over the public switched telephone network.

The SIVA database consists of four speaker categories: male users, female users, male impostors, female impostors. Speakers were contacted via mail before the test, and they were asked to read the information and the instructions provided carefully before making the call. About 500 speakers were recruited using a company specialized in selection of population samples. The others were volunteers contacted by the institute concerned.

Speakers accessed the recording system by calling a toll free number. An automatic answering system guided them through the three sessions that constituted a recording. In the first session, a list of 28 words (including digits and some commands) is recorded using a standard numbered prompt. The second session is a simple unidirectional dialogue (the caller answers prompted questions) where personal information is asked (name, age, etc.). In the third session, the speaker is asked to read a continuous passage of phonetically balanced text that resembles a short curriculum vitae.

The signal is a standard 8kHz sampled signal, coded using 8 bits mu-law format. The data collected so far consists of: MU: male users 20 speakers, 18 repetitions, FU: female users 20 speakers, 18 repetitions, MI: male impostors: 400 speakers, 1 repetition, FI: female impostors: 400 speakers, 1 repetition.

Price for ELRA members: R: 1,000 ECU/C: 3,000 ECU Price for non members: R: 3,000 ECU/C: 4,500 ECU

ELRA-S0054 Chilean Spanish FDB-500

This speech database gathers Spanish data as spoken in Chile. All participants are native speakers. The corpus consists of read speech, including digits and application words for teleservices, recorded through an ISDN card. There is a total of 507 speakers (272 male, 235 female). Each speaker pronounced a total of 24 utterances. The age class is divided as follows: 33 speakers are less than 16 year old, 215 speakers are between age 16 to 30, 207 speakers are between age 31 to 45, 51 speakers are between age 46 to 60, and 1 speaker is over 60.

The callers spoke 74 different items in total: isolated digits, yes/no, common application words.

The data is provided with orthographic transliteration for all 12,168 utterances including 4 categories of non-speech acoustic events. A phonetic lexicon with canonical transcription in SAMPA is also included.

The speech files are stored as sequences of 8 bits 8 kHz A-law samples. Data are stored in a SAM file format.

Price for ELRA members: 6,000 ECU Price for non members: 10,000 ECU

ELRA-S0059 ILE: Italian LExicon

ILE is a 588,000 entry Italian lexicon transcribed with SAMPA notation. It was generated, mainly for speech recognition purposes, by means of a morphological analyzer. Each stem was combined with all its possible suffixes to form valid words. Verbal forms do not include clitics.

The morpho-lexicon was obtained by properly processing an Italian dictionary, and adding by hand all possible inflections. This base lexicon was then enriched with names and neologisms found in the 65,000 most frequent words of the newspaper "Il Sole 24 Ore". Also, the most frequent Italian proper names and surnames (from the telephone directory), geographical names, acronyms, company names, commonly used foreign words were added to the lexicon.

All words are transcribed using SAMPA units for the Italian language. In case of multiple pronunciations for a word, one row for each different transcription is provided (a total of about 601,000 different transcriptions are provided for the 588,000 words lexicon). Stressed vowels are marked with the ASCII character ". Also, foreign words are transcribed using only SAMPA units for the Italian language, which leads to some awkward but effective transcription, at least for speech recognition purposes.

Price for ELRA Members: R: 3,000 ECU/C: 12,000 ECU Price for non Members: R: 6,000 ECU/C: 18,000 ECU

ELRA-S0060 MULTEXT Prosodic database

This database comprises one CD-ROM for each five languages (French, English, Italian, German and Spanish), totalling 4 hours and 20 minutes of speech and involving 50 different speakers (5 male and 5 female per language). The recordings on which the corpus is based consist of passages of about five sentences extracted from the EUROM.1 speech corpus.

The corpus was stylised automatically by an algorithm which factors out microprosodic effects and represents the intonation contour of utterances by a series of target points. Once interpolated by a smooth curve (spline), these points produce a contour indistinguishable from the original when re-synthesised, apart from a few detection errors. A symbolic coding of the 50,000 pitch movements of the corpus is also provided, along with the time-alignment of orthographic transcription to signal at word level. The entire corpus was verified and manually corrected by experts for each language.

The CD-ROMs contain for each passage:

The signal file from EUROM.1, the alignment of orthographic transcription to signal at word level, the Fo file, the stylisation files, the resynthesis using the stylised Fo, the symbolic coding file, the residual Fo, i.e. the difference between the Fo and the stylised curve, a description file for the recording.

Additional information: Campione, E., Véronis, J. (1998). A multilingual prosodic database. Proceedings of ICSLP'98, Sydney, Australia.

Price for ELRA members: R: 45 ECU/C: 2,000 ECU Price for non members: R: 100 ECU/C: 5,000 ECU

ELRA-S0061 French Speechdat(II) FDB-1000

This French telephone speech database is designed for development and assessment of French speech recognizers. It contains 48 utterances (40 mandatory and 8 optional items) for 1,017 different speakers, collected over the fixed telephone network. The database was produced by MATRA COMMUNICATION and was sponsored by the European Commission (CEC DGXIII), under the project LE2-4001. 17 speakers have been added to the original 1,000 speakers to fit the requirements of the database. The database complies with the common specifications designed in the SpeechDat(II) project. The main content of the database is speech and orthographic transcription files.

The speech files are stored as sequence of 8-bit, 8kHz A-law speech files and are not compressed, according to the specifications of SpeechDat. They contain a file header of 16 bytes. Each prompt utterance is stored within a separate file (file extension FRA) and has an accompanying ASCII SAM label file (file extension FRO).

Corpus contents: 5 application words; 1 sequence of 10 isolated digits; 4 connected digits: 1 sheet number (5+ digits), 1 telephone number (9-11 digits), 1 credit card number (14-16 digits), 1 PIN code (6 digits); 3 dates: 1 spontaneous date (e.g. birthday), 1 prompted date (word style), 1 relative and general date expression; 2 word spotting phrases using an application word (embedded); 1 isolated digit; 3 spelled-out words (letter sequences): 1 spontaneous, e.g. own forename; 1 spelling of directory assistance city name; 1 real/artificial name for coverage; 1 currency money amount; 1 natural number; 5 directory assistance names + 1 spelled-out name: 1 spontaneous, e.g. own forename, 1 city of birth / hometown (spontaneous); 1 most frequent city (out of 500); 1 most frequent company/agency (out of 500); 1 “forename surname”, 1 spelled-out city of birth; 2 questions, including "fuzzy" yes/no: 1 predominantly "yes" question, 1 predominantly "no" question; 9 phonetically rich sentences; 2 time phrases: 1 time of day (spontaneous), 1 time phrase (word style); 8 phonetically rich words.

Price for ELRA members: R: 9,000 ECU/C: 18,000 ECU Price for non members: R: 22,000 ECU/C: 25,000 ECU

Special discounted price when purchased with FRESCO (ELRA-S0016 or ELRA-S0017).
Special discounted price for those who purchased FRESCO after 1 October 1997.

ELRA-W0017 MULTEXT JOC Corpus

This CD-ROM contains a part of the corpus developed in the MULTEXT project financed by the European Commission (LRE 62-050). This part contains raw, tagged and aligned data from the Written Questions and Answers of the Official Journal of the European Community. The corpus contains approx. 5 million words in English, French, German, Italian and Spanish (approx. 1 million words per language). About 800,000 words were grammatically tagged and manually checked for English, French, Italian and Spanish, i.e. roughly 200,000 words per language. The same subset for French, German, Italian and Spanish was aligned to English at the sentence level.

The JOC corpus is delivered in Corpus Encoding Standard conformant format at each level of treatment :

- paragraph annotation level, conformant to the CESDOC specifications (1 M words * 5 languages);
- morpho-syntactic annotation level (PoS Tagging), conformant to CESANA specifications (200,000 words * 4 languages);
- parallel text alignment at sentence level, conformant to CESALIGN specifications (200,000 words * 4 languages).

Additional information: http://www.lpl.univ-aix.fr/projects/multext

Price for ELRA members: R: 45 ECU/C: 2,000 ECU Price for non members: R: 100 ECU/C: 5,000 ECU

ELRA-W0018 ARCADE/ROMANSEVAL corpus

The ARCADE/ROMANSEVAL corpus was used as a reference corpus in two international competitions:

- ARCADE, an exercise on multilingual text alignment financed by AUPELF-UREF
- ROMANSEVAL, part of the SENSEVAL exercise sponsored by ACL-SIGLEX and EURALEX, on word sense disambiguation.

The corpus contains raw data from the JOC corpus developed in the MULTEXT project financed by the European Commission (LRE 62-050), composed of 1 million words in English and four Romance languages: French, Italian, Spanish and Portuguese (Written Question and Answers from the Official Journal of the European Commission).

The annotation concerns all the contexts of 60 different test words (20 nouns, 20 adjectives, 20 verbs), i.e. ca. 3,700 contexts altogether, and comprises: semantic tagging of all the occurrences of the test words in the JOC corpus for French and Italian; word-level alignment of all the occurrences of the test words between French and English.

Additional information:http://www.lpl.univ-aix.fr/projects/arcade http://www.lpl.univ-aix.fr/projects/romanseval

Price for ELRA members: R: 45 ECU/C: 2,000 ECU Price for non members: R: 100 ECU/C: 5,000 ECU

ELRA-L0010 MULTEXT Lexicons

This CD-ROM contains a set of lexicons developed in the MULTEXT project financed by the European Commission (LRE 62-050). The set contains the following languages: English, French, German, Italian and Spanish.

English	66,214 Word forms	French	306,795 Word forms	German	233,861 Word forms
Italian	145,530 Word forms	Spanish	510,710 Word forms		

The MULTEXT lexicons are three-column tables, separated with a tabulation: the first column contains the word-form, the second column contains the lemma, and the third column contains the morpho-syntactic information associated to that form. This information is conformant with the MULTEXT/EAGLES specifications.

Additional information: http://www.lpl.univ-aix.fr/projects/multext

Price for ELRA members: R: 45 ECU/C: 2,000 ECU Price for non members: R: 100 ECU/C: 5,000 ECU

ELRA-T0362 NEWBASE (Extended version of ELRA-T0090 GEOBASE)

The terms were selected and collated by *Dr M.S.N. CARPENTER* during the course of his translation activities over the past ten years. The terms have been validated by publication in the scientific literature. Conceived as a bilingual terminological resource, it is also suitable for use in the development of translation memory systems.

Field types: administrative data; antonyms; abbreviations, contexts, cross-references, definitions; examples; grammatical label; local spelling variants; notes; sources; sub-domains; symbols; synonyms.

Main subject areas: GEOLOGY; structural geology; geochemistry; stratigraphy; sedimentology; geochronology; geophysics; seismology; physical geography; petrography; palaeontology; volcanology; marine geology; hydrogeology.

The database contains the following information:

- French part: 2940 French headwords, 2031 definitions in French (1275 terms have one definition at least), 175 contexts extracted from learned articles, 170 examples of specific terms related to main entry terms, 549 technical notes, observations or remarks, 733 close equivalents or other usages of main entry terms, 248 synonyms in French, 37 spelling variants in French, 40 antonyms in French, 760 cross-references to French entry terms.
- English part: 2965 English equivalents, 940 definitions in English (780 cards consist of one definition in French matching with one definition in English), 54 contexts, 132 examples, 221 technical notes, 1075 close equivalents or other usages of main entry terms, 307 synonyms in English, 128 local spelling variants, 35 antonyms, 573 cross-references to English main entry terms.

Total number of records: 479 cards extracted from cited bibliographic sources, 740 cards signed by person responsible (terminologist/trainee), and 2211 citations of bibliographic sources.

Specific terms related to the main entry term are tagged as examples. Polysemy is filtered according to grammar and/or usage in a particular sub-domain. Multiple translation equivalents in different sub-domains are each treated as an other form group (English target language).

Price for ELRA members: R: 3420 ECU/C: 4788 ECU Price for non members: R: 4788 ECU/C: 6840 ECU

ELRA-T0363 HYDROGEOLOGY DATABASE

400 terms

275 definitions in French

297 definitions in English

The terms were selected and collated by *Dr M.S.N. CARPENTER* during the course of his translation activities over the past ten years. The terms have been validated by publication in the scientific literature. Conceived as a bilingual terminological resource, it is also suitable for use in the development of translation memory systems.

French-English hydrogeology terminology extracted from "Le forage d'eau - réalisation, entretien, réhabilitation", Michel DETAY, pp. 379, Masson, Paris, 1993, compiled following translation into English by *Dr M.S.N. CARPENTER* (“Water wells: implementation, maintenance and restoration”, 1996, coeditor John Wiley, Chichester, U.K.).

Subject areas include: groundwater hydraulics, hydrology, water chemistry.

Price for ELRA members: R: 850 ECU/C: 1190 ECU Price for non members: R: 1190 ECU/C: 1700 ECU

ELRA-T0364 PEDOLOGY DATABASE

453 terms

358 definitions in French

143 definitions in English

The terms were selected and collated by *Dr M.S.N. CARPENTER* during the course of his translation activities over the past ten years. The terms have been validated by publication in the scientific literature. Conceived as a bilingual terminological resource, it is also suitable for use in the development of translation memory systems.

French-English pedology terminology, extracted from an INRA/CILF document and other sources (TERMIUM, Concise Oxford Dictionary of Earth Sciences, etc.). Records compiled using index file (from Mme BOUROCHE, INRA, corrections delivered 15/04/96) and then merged with trainee project work (POUIVE) to form database TERMSOL.XM8.

Subject areas include: soil science, soil mechanics, geomorphology, geology, physical geography, meteorology, hydrology, hydrography, mineralogy.

Price for ELRA members:R: 585 ECU/C: 819 ECU Price for non members: R: 819 ECU/C: 1170 ECU