The ELRA Newsletter

EUROPEAN LANGUAGE RESOURCES ASSOCIATION

June 1997

Vol.2 n.2

Table of contents

The ELRA Catalogue, June 1997, Release 2.2 is enclosed

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:
Achim Blatt
Louis Boves
Laila Dybkjær
Adam Kilgariff
Paulo Martins
Niels Ole Bernsen

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@calvanet.calvacom.fr
WWW:
http://www.icp.grenet.fr/
ELRA/home.html

Dear ELRA Members,

This edition of the ELRA Newsletter offers updates and report on activities by members of all three ELRA colleges and others, as well as by ELRA itself. In addition to information on Termcat (the Catalan terminology centre) and the CEC's Euramis project, we are offering an overview of speech research activities and corpora in the Netherlands. On the research side, we have an article on evaluating word sense disambiguation and a progress report on DISC (the ESPRIT Concerted Action on Spoken Language Dialogue Systems and Components).

One ELRA-related article is an overview of our planned market study, designed to define and identify the market structure and ascertain user needs and expectations. This is one of the main focuses of our work at the moment; in an initial step, the sector has been broken down into segments and a list of action items drawn up for each one. Our second market-related activity is a fundamental redesign of the Catalogue of Resources, in order to provide more details and samples. Our thanks go to those members and others who contributed to the user needs survey or were interviewed by us, and we would like to ask for your continued support and input for the activities mentioned above.

In this context, we would like to welcome Malin Nilsson, who has now started work at ELRA/ELDA as a marketing assistant. Another marketing innovation is that non-Europeans now also have the opportunity to subscribe to ELRA - the appropriate forms can be obtained from our Web site.

Of course, work on ELRA's basic tasks - the collection and dissemination of resources - has also continued during the period since the last News, and a list of new resources with complete descriptions is to be found on the last page of this newsletter. A number of new projects are also in the pipeline; more details will follow in the next issues. Considerable work has been put into enhancing and extending internal systems procedures, and hence the services to our members as well. Last but not least, plans for the future include an international conference on Language Resources and Validation, to be held in co-operation with leading language engineering associations and institutions throughout the world. This event, the first of its kind, is tentatively scheduled for May 1998.

We hope that this short overview gives an idea of how much progress has been made at ELRA over the past few months. In this context, we would like to invite all those members who have not yet renewed their subscriptions to do so - and to take advantage of our special offer for free resources when they do so.

Members can choose up to two of the following resources: the TED translingual English database (speech); the MLCC parallel written and multilingual corpora (9 and 6 languages respectively) and the CRATER parallel written corpus (telecommunication, 3 languages); and a multilingual terminology database containing over 20,000 terms in a number of domains including finance, telecommunications, energy and the environment

With best wishes,

Antonio Zampolli, President

Khalid Choukri, CEO

ELRA Profiles

Malin Nilsson, ELDA Marketing Assistant

Born in 1971 in Halmstad, Sweden, Malin Nilsson studied economics, business management and statistics at the University of Gothenburg, Sweden and the Plymouth Business School in the United Kingdom before obtaining a Masters in Business Administration from the University of Gothenburg's Business School in 1996.

After finishing university, she worked on a marketing project in Tartu, Estonia for the Swedish adult training school Folkuniversitetet, which included surveying supply and demand for different training activities on the Estonian corporate training market. At ELRA/ELDA, Malin will contribute to the upcoming market survey project as well as handling other marketing and promotional issues for the organisation.

EURAMIS : The European Advanced Multilingual Information System

Achim Blatt and Paulo Martins

Introduction

Although desktop computers are becoming increasingly powerful in terms of both storage capacity and processing, it will not be possible in the foreseeable future to store all useful linguistic data, or to run all useful programs in the domain of natural language processing locally. In addition, some programs are just too big to run on the desktop - or they have to look through such large amounts of data that they would be unacceptably slow if they did so. This means that if linguistic data and applications are to be made available to all potential users, they must be available centrally - whereas data which are only of interest to a limited number of users should be kept locally.

EURAMIS was designed taking into account both the considerations above and the size of the European Commission's Translation Service, one of the largest translation services in the world (1 million pages are translated per year). The system is based on a client/server architecture communicating via E-mail, with a single common interface on the client side which enables the user to launch different requests from the same environment. The server is able to channel the requests to the various applications, combining different services and thus achieving synergy effects between them. The applications are built in a modular way and use a common pivot format (SGML, Unicode), so that the system is independent of proprietary formats.

The project was launched in 1994 by the European Commission with a call for tenders for the "Development of multilingual tools and their integration into multilingual services". It has been under development since the beginning of 1995 and is currently being put into production, with a gradual growth in the number of users until eventually it becomes accessible to the whole Commission Translation Service. It is planned that, at a later stage, some products will also be made available to other services that create documents.

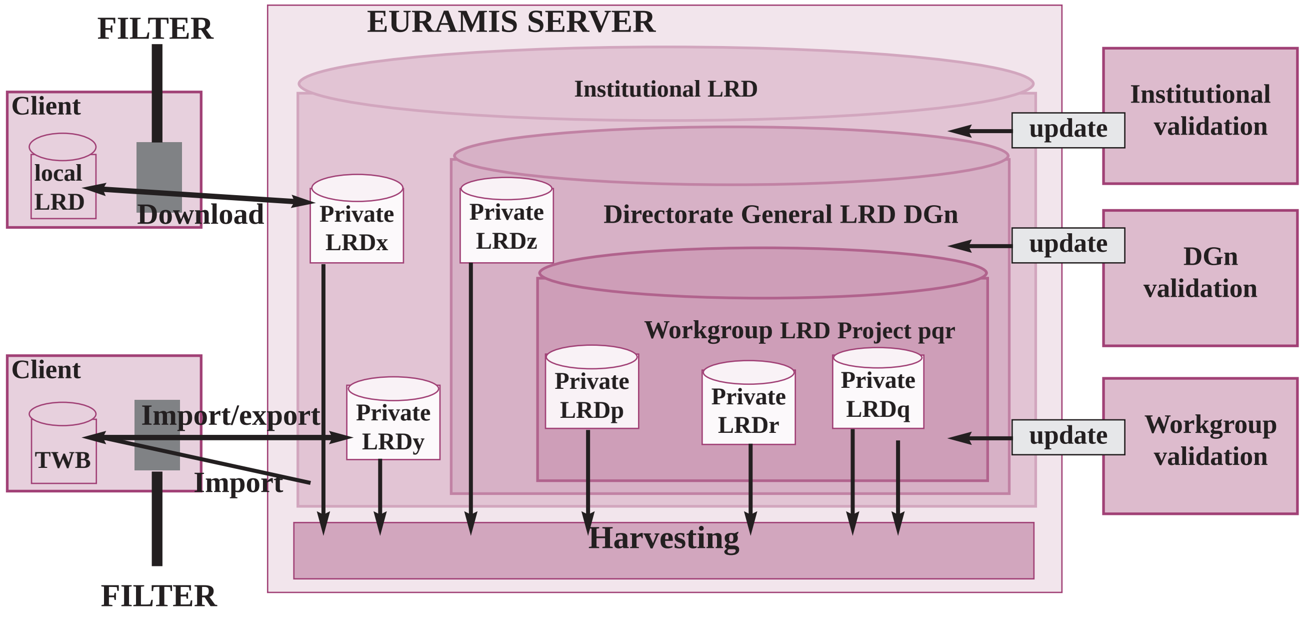
The purpose of the project is to offer a complete set of linguistic tools that can be used in the translation process. Some of these tools were developed in the-course of the project, others existed already. EURAMIS was designed to integrate all these tools in a single and coherent working environment, creating the mechanisms necessary for these tools to communicate with each other.

The database

The LRD - Linguistic Resources Database - is the storage area for all the EURAMIS linguistic data, which comes from various information sources. It is made up of databases for individual use, work group databases and a central database accessible by everyone. The LRD incorporates:

- translation memories: aligned sentences with references to the source document;
- central terminology and phraseology, mainly from EURODICAUTOM but also from other sources;
- in the future, machine translation dictionaries will be included, with syntactic and morphological information which will also be accessible by other applications.

Access to the various databases is arranged hierarchically, allowing the user to choose the order in which the system will look for data. Procedures have been designed for "harvesting" the individual databases and for their consolidation into bigger databases. In turn, these can be integrated into an even larger central database as shown here:



The products

To reflect the integrated nature of the different products within the project, a unique interface was built to access all the services. This EURAMIS Client Interface is the entry point for use of the services developed under the EURAMIS project and also to a number of already existing linguistic products. It is through this interface that the user specifies the parameters necessary for each request and transmits them to the server.

Machine translation: the user can ask for a translation to be provided for any of the 16 language pairs available in the Commission's system, with four source languages: English, French, Spanish, and German.

CELEX titles look-up: the CELEX database (Communitatis Lex) contains all European legislative texts. This service provides the titles of CELEX documents which are referred to in a submitted document.

Terminology: the central terminology database of the Commission, EURODICAUTOM, is accessible in batch mode via the EURAMIS interface. This service provides an automatic EURODICAUTOM look-up for all single- or multi-word terms which can be found in a given document. This is done in a two-step procedure: identification of terms with the Commission's machine translation system, and look-up in EURODICAUTOM proper.

Translation memory tool set: the translation memory tool set comprises several applications which carry out two main tasks: sentence alignment and storage of the results in the databases; and translation memory retrieval.

For the first task, a very powerful alignment application was developed, which produces very good results due to a high degree of customisation. It basically uses well known statistical methods, which will be complemented in future by anchors (styles, numbers, acronyms, etc.). For the human correction of misalignments, a Windows **Alignment Editor** was developed. Each alignment made is stored in the database in multilingually aligned structures: although the alignment is always made between two languages, database structures are multilingual. In other words, if the source sentence of a new alignment already exists in the database and if the secondary information (domain, text type, requesting service) allows it, the system stores the target sentence in the same structure.

For translation memory retrieval, the key aspect is performance of the search phase. This is done at sentence level and includes a "fuzzy" component, which is based on trigram technology. Fuzziness is reduced by automatic replacement of repetitive elements like months, years, numbers, etc.

Users can:

- specify the maximum degree of fuzziness they accepts;
- indicate a sequence of translation memories for their query (e.g. their own, their group's, etc.);
- narrow down the search to specific domains, requesting departments, text types or similar features, or explicitly exclude a given domain, etc. from search.

The results can be shown in three ways:

- in a word processing file, in which case only one result can be given for each sentence;
- in the export format for the Trados Translator’s WorkBench (TWB), which is used as a local translation memory tool;
- in the EURAMIS format, to be viewed in the EURAMIS Text Editor, a specific editor developed in the project.

Integration of translation memory and machine translation: for the sentences where no satisfactory match is found in the translation memory according to the parameters indicated, the system provides a machine translation result.

Text analysis: The prime aim of automatic text analysis is to extract potential terminology and phraseology by conducting a repetition analysis within and across documents. The user can provide a large number of parameters, such as the minimum length of potential expressions, a list of expressions to ignore or a list of stop words, with the possibility of specifying their position (before, within, or after potential terms).

The products mentioned below are still under development and the production phase will only start at the beginning of next year.

Replacement below sentence level: this application is based on a PC application which is widely used in the Translation Service. It creates a mixture of source text and translation where the most repetitive parts are already replaced by target expressions (single words or sequences of words up to a sentence long) by comparing the expressions in the databases indicated with the text to be translated.

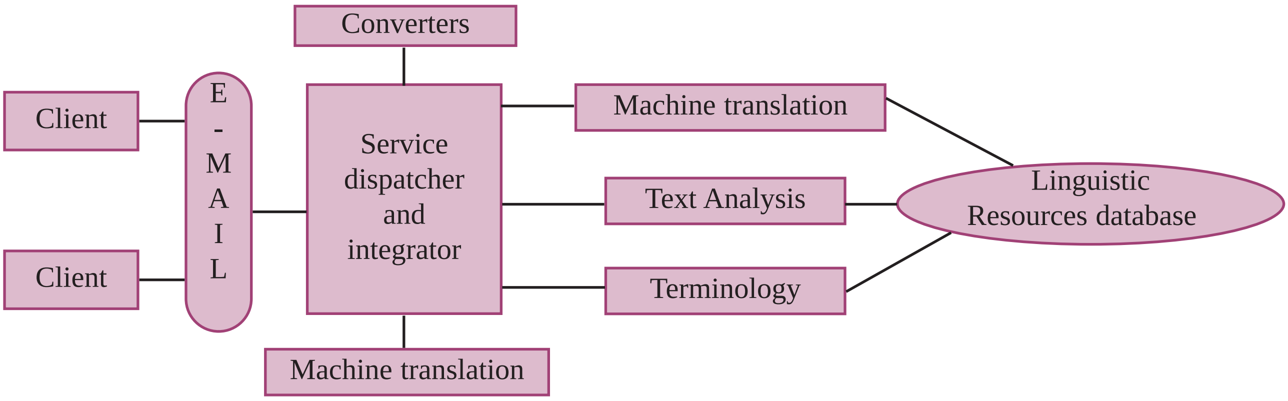
Content identification: this tool is used for the identification of the type of text (letter, contract, legislation, etc.) and the language(s) used in it. The system searches in the text for words that are unique to a language or a text type and are listed in a specific file, taking into account their position and frequency. This tool can be integrated in the translation memory process, for example to mark parts of the text where the language differs from the one indicated by the user, or it can be used in order to classify large numbers of documents. Currently, an extension of this approach using trigram technology is being considered.

Combination of services: the architecture of the EURAMIS project (see below) makes it possible to combine different standalone modules in order to create new, interesting services with little additional effort. One example, which will be built almost entirely using existing components, is the automatic creation of an ad hoc translation memory on the basis of CELEX documents that are referenced in a given text. To do that, the system:

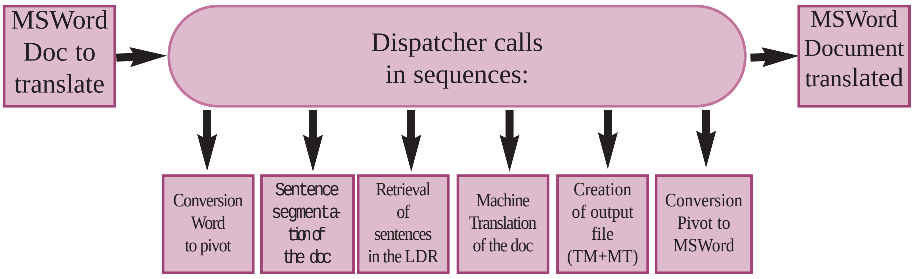
- calculates the CELEX identifier that corresponds to each legislative document that is referenced in the text;
- retrieves the respective CELEX documents in the source and target languages and aligns them (the automatic alignment of these texts is in general very accurate);
- creates a document-related translation memory from these alignments.

Client/server architecture

The whole project is based on a client/server architecture: the applications run on a very powerful Unix server (other related external applications, like machine translation, run on mainframes). After the user creates the request using the Windows EURAMIS client interface, it is submitted to the server via e-mail (X.400), where the applications needed are sequenced by a dispatcher. The results are sent back to the user’s personal mailbox. The following diagram illustrates the general architecture:



The service dispatcher is the core of the messaging system, allowing the user to combine services. If, for instance, the translator wants a text to be submitted to translation memory and then to machine translation, he or she will request this combination as one service, and the dispatcher will call in sequence the applications involved to fulfil the request:


MSWord Doc to translate
Dispatcher calls in sequences:
MSWord Document translated
Conversion Word to pivot
Sentence segmentation of the doc
Retrieval of sentences in the LDR
Machine Translation of the doc
Creation of output file (TM+MT)
Conversion Pivot to MSWord

To achieve this level of modularity and integration, the EURAMIS applications use a powerful pivot format for the data, with specific interfaces. This pivot format (SGML and Unicode) is used wherever possible. Nevertheless, each request may be submitted to a number of format conversions, since the system acts like an integrator, receiving different document formats (WordPerfect, Word, etc.) and communicating with several external applications (EURODICAUTOM, machine translation, Translator's WorkBench).

For more information, please contact:
Achim Blatt,
E-mail: Achim.Blatt@sdt.cec.be
Paulo Martins,
E-mail: Paulo-Martins@sdt.cec.be
Translation Service
European Commission
L-2920 Luxembourg

Speech Research Activities and Corpora in the Netherlands

Louis Boves

This paper gives a short overview of the major ongoing speech technology research and development activities in the Netherlands. The emphasis is on projects which lead to the creation of speech corpora. In passing, several activities in Flanders, the Dutch speaking part of Belgium, are also mentioned.

The most important players in the speech technology field in the Netherlands are KPN (Royal Dutch PTT), the Institute for Perception Research in Eindhoven (this used to be a joint venture of Philips Research and the Technical University, but was turned into a university institute as of March 1, 1997), and Nijmegen University. Other universities, mainly those in Amsterdam, Leyden and Utrecht, have smaller efforts in the field, under the umbrella of the Foundation for Speech Technology. This Foundation is also the parent organisation of the Speech Processing Expertise Centre (SPEX), a non-profit organisation which aims to create and validate spoken language resources and to make these available to the academic and commercial speech R&D in the country. After having profited from government subsidies during its first years, SPEX has been self-supporting since the beginning of 1996. Among other things, it has produced the Groningen Corpus (read speech), the Speech Styles Corpus (read speech, picture descriptions and spontaneous speech by the same speakers) and the Dutch Polyphone Corpus (5,000 speakers, uniformly distributed across the two sexes and all regions of the country). In addition, SPEX has collected several corpora of speech coming from the cellular networks. These corpora were commissioned by KPN (both KPN Research and commercial business units in PTT Telecom). The availability of the cellular corpora must be negotiated with KPN.

In Flanders, the most important players are Lernout and Hauspie Speech Products, the University of Gent and the University of Leuven. In 1995 and 1996, a two-year government-funded research programme in the field of Speech and Language Technology was carried out, mainly by the two universities mentioned above, plus a contribution by Antwerp University. These efforts have resulted in the addition of Flemish pronunciation variants to the CELEX lexical corpus for Dutch (the FONILEX corpus), several hours of transcribed speech recorded off the public radio and a medium-sized corpus of read and spontaneous speech recorded in quiet environments with wide band equipment.

The Dutch Science Council, NWO, is probably the most important funding agency for speech technology research. NWO funds the national programme 'Language and Speech Technology', in which four universities (Nijmegen, Eindhoven/IPO, Amsterdam and Groningen, the latter two exclusively working in natural language processing) collaborate with Philips Research and KPN Research. Although this programme is aimed at basic scientific research in speech recognition, speech generation, NLP and dialogue management, we still intend to build a working prototype of an advanced spoken dialogue system (for the time being in the domain of public transport information). We consider an operational prototype system to be the only convincing proof that newly developed technology really outperforms existing modules.

The NWO Programme has obvious links to the LE project ARISE. In addition, it is closely linked to a commercial project by PTT Telecom, Philips Dialogue Systems and Openbaar Vervoer Reisinformatie (OVR, a company that provides a public transport information service via a single nation-wide premium-rate telephone number). To reduce the number of calls abandoned in the waiting queue, OVR intends to introduce the Philips Automatic Information System in the course of 1997, to handle at least part of the callers who only need information on train schedules. The automatic service will be installed and maintained by PTT Telecom, and leased by OVR.

As preparation for the commercial roll-out, a total of over 10,000 human-machine dialogues will be recorded and orthographically transcribed in order to make the material suitable for training the speech recogniser and the NLP module that interprets the utterances. Since the vocabulary of the application domain is rather limited (in the first 5,000 dialogues no more than c. 1,500 different words were observed) the material should also make an excellent tool for investigating variations in pronunciation. The corpus, the IPR rights of which will reside with PTT Telecom, will be available to the NWO Programme for research purposes. An attempt will also be made to make it available to the research community at large.

Within the framework of the NWO Programme, part of the corpus will be annotated at the syntactic and semantic levels, so as to make it suitable for training advanced probabilistic NLP systems. Equally, 5,000 dialogues between arbitrary callers and human operators have been transcribed on behalf of OVR. The availability of this corpus should be discussed with OVR.

KPN and Nijmegen University are participating in the LE project CAVE. A large part of the technological R&D work in CAVE is based on the so called SESP corpus collected by SPEX on behalf of KPN Research in 1995. 50 speakers (half of whom were females) called 25 times from all kinds of locations and from all over the world and spoke 14-digit card numbers, 4-digit PIN codes, and several spontaneous answers to questions. The corpus has recently been extended to include over 100 speakers, although most of the additional speakers called only ten times, and only from the domestic network. In the course of 1997, the extended SESP corpus is will be made available for research purposes.

In 1997, a large corpus collection project will start, funded by the Dutch and Flemish governments. The overall design of this corpus will resemble that of the well known British National Corpus (BNC). It will contain approximately 10 million words, half of which will be more or less formal speech (sermons, lectures, discussions during meetings, interviews, etc.), while the other half will consist of surreptitiously recorded conversations of subjects who will be recruited so as to cover both the main regional variants of the language, and three levels of social status and education. The complete corpus will be transcribed orthographically and annotated at the lexicological and syntactic level. Part of the corpus will also be enriched with detailed phonetic and prosodic annotations and semantic codings. The full corpus is scheduled to be available in the year 2000; partial versions (less words, only basic transcription and annotation) will be available as of the beginning of 1998. Unlike the BNC, the Dutch corpus will come as a combination of speech signals and transcriptions/annotations. For the complete corpus, a direct link between transcription data and the acoustic signal will be provided. This corpus is meant to take speech and language research a significant step further than can be achieved with the mainly read speech in the Polyphone-like corpora. Of course, the much more complex types of speech in the Dutch corpus will necessitate the development of advanced transcription and annotation techniques. This corpus project will be led by SPEX and CELEX. the lexical Expertise Centre.

Last but not least, Cito (the Central Institute for the Development of Educational Tests), Swets Testing Services and PTT Telecom and Nijmegen University have started a project aiming to develop automated testing of the pronunciation quality of adults who learn Dutch as a second language. A spin-off of this project will be a large corpus of read and spontaneous speech produced by speakers with a very wide range of native languages.

For more information, please contact:
Louis Boves
SPEX (Centre for Speech Processing
Expertise)
Postbus 421
2260 AK Leidschendam The Netherlands
Phone: +31 24 361 29 02
Fax.: +31 24 361 59 39

The DISC Project

Niels Ole Bernsen and Laila Dybkjær

Introduction

Based on a recent presentation at the SALT Workshop (Bernsen and Dybkjær 1997), this paper briefly presents the aims and assumptions of DISC, the Esprit Long-Term Research Concerted Action No. 24823 "Spoken Language Dialogue Systems and Components: Best Practice in Development and Evaluation," which started on 1 June 1997. The DISC partners are the Maersk Mc-Kinney Moller Institute for Production Technology (MIP), Odense University, Denmark (co-ordinator); Human-Machine Communication Department, Centre National de la Recherche Scientifique (CNRS-LIMSI), France; Institut für Maschinelle Sprachverarbeitung (IMS), Universität Stuttgart, Germany; Department of Speech, Music and Hearing, Kungliga Tekniska Högskolan (KTH), Sweden; Vocalis Ltd, UK; Daimler-Benz, Germany; and Stichting Elsnet, The Netherlands.

The need for best practice in development and evaluation

No current scheme tailors software engineering best practice to the particular purposes of dialogue engineering, i.e. to the development and evaluation of spoken language dialogue systems (SLDSs). The goals of dialogue engineering include optimisation of the user-friendliness of SLDSs, a factor that will ultimately determine their rank among emerging input/output technologies. DISC aims to develop the first detailed and integrated set of development and evaluation methods and procedures (guidelines, checklists, heuristics) for dialogue engineering best practice, as well as a range of support concepts and software tools. The methodology developed by DISC will contribute toward establishing dialogue engineering as a sub-discipline of software engineering.

At this time neither accepted standards (or even widely understood benchmarks) for assuring potential SLDS customers or users of the quality of the systems exist, nor are there any reliable methods for comparing the quality of two SLDSs before selecting one for deployment in the field. In an increasingly competitive marketplace, the ability to state that a system has been developed according to a carefully designed and validated dialogue engineering methodology, along with the ability to report evaluation results in a standardised framework, are likely to give products developed in this way a competitive advantage. This in turn may stimulate the take-up of the methodology by other organisations.

SLDS technology is taking off on a broad scale. Current estimates are that the global annual market for speech recognition alone will be $8 billion in the year 2000. According to the Ovum report on voice processing published last year, the global voice processing market in 1996 amounted to $2.1 billion and was expected to grow to $2.9 billion in 1997 and $3.75 billion in 1998. Even if -by a conservative estimate - only 1% of this field could be identified as SLDSs, that is still a very large figure. The bulk of this business is in the US, but the opportunity to take a large and increasing share remains open to Europe.

Current commercial SLDSs are able to carry out routine tasks that were previously done by humans, thus generating significant savings for the companies or institutions that install these systems. During the last few years, interactive speech technology has begun to enjoy significant deployment in real-world applications in large vertical markets such as banking, finance and market research, as well as in telecommunications. An upcoming domain for advanced SLDSs is that of train information. Perhaps the most advanced SLDS in commercial use has been developed by Philips and is used by Swiss Rail. The system is based on the Philips Automatic Train Timetable Information System, a demonstration model of which has been publicly available since February 1994 in Germany by calling +49 241 604020. The development of similar train information systems is underway in the Netherlands, France and Italy. More advanced and flexible large vocabulary SLDSs and systems integrating speech into multimodal systems are progressing from research laboratories to industrial exploitation and will have commercial significance by the end of the DISC action.

Publicly funded research has provided the major driving force for the technological advances exemplified by these systems. In the US, this research has been co-ordinated by DARPA (previously called ARPA) through its competitive evaluations in large vocabulary speech recognition (Resource Management task) and spoken language understanding (ATIS task). The special issues associated with spoken language dialogue have been more clearly addressed in Europe than in the USA. Projects such as SUNDIAL, the Danish project on Spoken Dialogue Systems, MAIS, RAILTEL and VERBMOBIL have established a strong base of SLDS expertise in Europe. The results achieved by several of those projects are currently being taken a step further towards commercialisation in the LE ARISE project.

Despite unquestionable progress - particularly in those parts of the SLDS components field that have been delivering commercial applications for more than a decade - the design, development and evaluation of usable SLDSs are today as much of an art and a craft as they are an exact science with established standards and procedures of good engineering practice. The route from the initial idea through analysis, requirements specification, design and evaluation cycles, prototype development, in-house and field testing to the final product and its evaluation is replete with unknowns and development steps that are undersupported in terms of procedures, concepts, theory, methods and software tools. Given the proven potential of SLDS technologies, there is a need to take a significant step forward by creating a best practice methodology for SLDS development and evaluation, and to start developing the concepts, methods and software tools required to integrate SLDS development into mainstream software engineering. DISC aims to make central contributions to an SLDS development and evaluation best practice methodology including novel concepts, methods and software tools.

DISC objectives, approach and envisaged results

As a long-term research Concerted Action aimed at making innovations responding to industrial needs, DISC aims to expand the state of the art in dialogue engineering in the following four ways:

(1) generalising current knowledge by utilising state-of-the-art expertise to analyse a broad range of current SLDS and components development and evaluation practices, thereby creating a detailed overview of current practice;

(2) maturing promising novel concepts, methods and software tools which exist in preliminary versions at partner sites and developing them to the industrial transfer stage, when possible;

(3) testing a comprehensive scheme of dialogue engineering best practice on industrial and research cases, to the extent possible within the duration of the Action;

(4) systematising results into a detailed, procedural dialogue engineering best practice methodology that takes a balanced view of competing approaches and technologies within current best practice, where such exist. The methodology should enable the user to specify the required behaviour (functionality, performance, ergonomics) and to determine the extent to which the system, its components and their interaction meet the stated requirements. Input from industry will be integrated into the methodology.

DISC will achieve its stated goals via three main work packages addressing current practice, best practice, and novel concepts, guidelines and software tools, respectively. Each work package will focus on a set of aspects of SLDSs, including speech recognition, speech generation, language understanding and generation, dialogue management, human factors and systems integration.

To ensure common approaches to each of the main results-building activities and to ensure that the results produced are compatible across different aspects, each approach will have to include a set of agreed evaluation criteria. Thus, (a) the common approach for mapping out current practice includes criteria for the evaluation of current practice; (b) the common approach for testing best practice methods and procedures on industrial cases includes criteria for evaluating the transferability of these methods or procedures; and (c) the common approach for iterative development and testing of novel concepts, methods and software tools includes criteria for determining the feasibility of specific development and testing projects, as well as for evaluating transferability. The approaches and criteria in (a) through (c) will form the basis for the quality assurance that can be achieved with the methods, procedures, concepts and software tools developed by DISC.

All partners are providing the Action with access to products and running prototypes and their components, as well as to prototypes under development. DISC will take advantage of existing practices, theories and tools, including the results of the US ARPA exercise in comparative SLDS evaluation; results emerging from the LRE EAGLES project in the fields of de facto standards and guidelines for speech products, natural language components and evaluation; and experience from national initiatives in component evaluation methodologies, such as the German Morpholympics, the French

GRACE project and other evaluation actions of the AUPELF group.

The industrial benefits of DISC will be the following:

- Progress towards the integration of SLDS best practice into software engineering.
- Improved feasibility assurance for development projects (risk minimisation) and more exact feasibility assessment.
- Improved procedures, methods, concepts and software tools.
- Reduced development costs and time; improved maintenance and reusability.
- Improved product quality and increased flexibility and adaptability.
- Progress toward the establishment of dialogue engineering standards.
- Improved guarantees to end users that a product has been developed following best software and cognitive engineering practice.

This will enable end users to assess objectively different systems and components technologies against one another and choose the right product according to quality, price and purpose.

The industries involved in DISC share the view that a best practice dialogue engineering methodology consisting of detailed procedures and methods, concepts, and software tools for development and evaluation will constitute an obvious competitive parameter for the emerging European SLDS supplier and end-user sector.

References

ARPA. Proceedings of the Speech and Natural Language Workshop. San Mateo, CA: Morgan Kaufmann, 1994.

Aust, H., Oerder, M., Seide, F. and Steinbiss, V.: The Philips Automatic Train Timetable Information System. Speech Communication 17, 249-262, 1995.

Aust, H. and Oerder M.: Dialogue Control in Automatic Inquiry Systems. Proceedings of the ESCA Workshop on Spoken Dialogue Systems, 121-124, Aalborg, Denmark, 1995. Also in: Proceedings of TWLT9, 45-49, Enschede, The Netherlands.

Bernsen, N.O. and Dybkjær, L.: The DISC Concerted Action. In R. Gaizauskas (Ed.): Proceed-ings of the SALT Club Workshop on Evaluation in Speech and Language Technology, Sheffield, June 1997, 35-42.

Bernsen, N.O., Dybkjær, H. and Dybkjær, L.: Designing Interactive Speech Systems. From First Ideas to User Testing. To be published by Springer Verlag, 1997.

Blyth, B. and Piper, H.: Speech recognition: a new dimension in survey research. Journal of the Market Research Society 36(3), 1994, 183-203.

DARPA. Proceedings of the Speech and Natural Language Workshop. San Mateo, CA: Morgan Kaufmann, 1992.

Dybkjær, H., Dybkjær, L. and Bernsen, N.O.: Design, formalisation and evaluation of spoken language dialogue. Proceedings of the TWLT9 Workshop, Enschede, 1995, 67-82.

Fraser, N.M. and Thornton, J.H.S.: VOCALIST: a robust, portable spoken language dialogue system for telephone applications. Eurospeech'95, Madrid, 1995, 1947-50.

Lamel, L., Bennacef, S., Bonneau-Maynard, H., Rosset, S. and Gauvain, J.L.: Recent developments in spoken language systems for information retrieval, Proceedings of the ESCA Workshop on Spoken Dialogue Systems, Vigsø, Denmark, 1995, 17-20.

Peckham, J.: A new generation of spoken dialogue systems: results and lessons from the SUN-DIAL project. Eurospeech'93, Berlin, 1993, 33-40.

Peckham, J. and Fraser, N.M.: Spoken language dialogue over the telephone. In H. Niemann, R. de Mori and G. Hanrieder (Eds.): Progress and Prospects of Speech Research and Technology. Sankt Augustin: Infix, 1994, 192-203.

Peckham, J. and Fraser, N.M.: Speech Understanding and Dialogue. Cambridge, MA: MIT Press, (forthcoming).

Wahlster, W.: Verbmobil - Translation of Face to Face Dialogues. Machine Translation Summary IV, Kobe, 1993.

Young, S.: Speech recognition evaluation: A review of the ARPA CSR Programme. In R. Gai-zauskas (Ed.): Proceedings of the SALT Club Workshop on Evaluation in Speech and Language Technology, Sheffield, June 1997, 197-205.

Niels Ole Bernsen and Laila Dybkjær
The Maersk Mc-Kinney Moller Institute for Production Technology
Odense University,
Campusvej 55,
5230 Odense M,
Denmark
emails:nob@mip.ou.dk, laila@mip.ou.dk
Phone: (+45) 65 57 35 44
fax.: (+45) 66 15 76 97

Termcat, the Catalan Terminology Center

Termcat was established in 1985 on the initiative of the Catalan government's General Directorate of Linguistic Policies. Designed as a centre for the co-ordination, creation, and dissemination of terminological resources, its aim is to guarantee the availability of language resources and to facilitate the use of Catalan in scientific and technical domains. It is supported by the Ministry of Culture of the Generalitat de Catalunya, the Institut d'Estudis Catalans, and the Consortium for Linguistic Normalisation.

Termcat's activities mainly focus on the domain of terminological research, i.e. the creation of resources, the organisation and management of research projects, methodological support, etc. Its objective is to fulfil the terminological needs of a number of different sectors.

Termcat has been involved in more than 250 terminological projects to date, including multilingual specialised dictionaries such as the *Diccionari de cartografia*, the *Diccionari de maquinària agrícola*, the *Diccionari de l'empresa elèctrica, ferroviària terminology*, the *Diccionari d'anatomia*, the *Diccionari de sociologia*, and the *Diccionari de lingüística*, as well as vocabularies and more widely disseminated lexica such as the *Vocabulari dels electrodomèstics*, the *lèxic de futbol americà*, and the *Lèxic de material d'oficina*.

Within the framework of the policy laid down by the Catalan Government's Linguistic Department, Termcat sets up terminology projects to provide the various socio-economic sectors with the specialised lexica they need. Among these are the *Diccionari visual de la construcció* (the first illustrated terminological dictionary in Catalan), the *Diccionari de bombers*, the *Diccionari policial*, and a collection of bilingual lexica for the industrial sector. In addition, it has revised the terminology of some statistical classifications such as, among other things, the *Classificació internacional uniforme d'ocupacions*.

In addition to promoting the systematic use of existing specialised lexica, Termcat's terminological activities are designed to help coordinate the creation of new terms and disseminate terminology among the users involved. To this end, Termcat is involved in the standardisation of Catalan terms, which means studying, proposing, and promoting neologisms in a manner which is not only in line with linguistic standardisation, but is also accepted by the specialists who will use them, and in harmony with international use.

Since the organisation was founded, Termcat's documentation service has responded to around 12,000 inquiries a year regarding the various terminological difficulties facing translators and writers of specialised texts. The increasing production of texts in Catalan has led to the articulation of a global service that integrates documentation information, text revision, organisation, and support for translation teams.

In order to guarantee the future of Catalan as a working language in the fields of science and technology, Termcat aims to ensure that Catalan terminology exists for new fields and technologies. The multilingual terminology work for the Atlanta Olympic Games Organising Committee, the preparation of the *Diccionari general tècnic* and the *Diccionari general social* for the METAL/X machine translation system, the *Diccionari d'assegurances*, the *electronic Vocabulari informàtic de la construcció* CONSTRULEX, and the participation in European projects of language engineering are all strong signs of this involvement.

For more information, please contact:
Termcat
Diputacio 119, 5a
08015 Barcelona - Spain
Tel.: +34 1 93 451 47 77
Fax.: +34 1 93 451 64 37
http://www.cinet.fcr.es/cs/termcat/

Evaluating word sense disambiguation programs

Adam Kilgarriff

The problem

Open a dictionary at random, choose a word at random the odds are, the dictionary says it has more than one meaning. When a word is used in a book or in conversation, though, generally just one of those meanings will apply. People don't have a problem. We are very rarely slowed down in our comprehension by the need to work out which meaning of a word applies. But computers are. The clearest case is in machine translation. If English *drug* translates into French as either *drogue* or *medicament*, then an English-French MT system needs to disambiguate *drug* if it is to make the correct translation.

People use the surrounding context to select the appropriate meaning. The context can be grammatical (if modified by a proper name, as in *AIDS drug*, it is probably medicament), lexical (if followed by *addict*, *trafficker*, or *squad* it is *drogue*), or domain-based (if the text is about policing, probably *drogue*, if about hospitals, probably *medicament*). People can usually disambiguate on the basis of very little surrounding context, with five words usually proving sufficient.

WSD programs

For thirty years now, people have been writing programs so that computers can do the same. This task is called Word Sense Disambiguation (WSD). Early programs required human experts to write sets of disambiguation rules for each multi-sense word. This was a problem, since it involved a huge amount of labour to write rule-sets or "Word Experts" for a substantial amount of the vocabulary.

The WSD problem can be divided into two parts. The first part is how do you express what meaning number 1 and meaning number 2 of a word are to the computer? The second is, how do you work out which of those meanings matches an occurrence of a word to be disambiguated? Lesk (lesk, 1986) took a novel tack, using the text of dictionary definitions as an off-the-shelf answer to the first problem. He then measured the overlap, in terms of words in common, between each of the definition texts and the context of the word to be disambiguated. Much recent work uses sophisticated variants of this idea.

Dictionary-based approaches remain tied to a particular dictionary, with concomitant errors, imperfections and copyright constraints. With the advent of huge computer corpora and computers powerful enough to compute complex functions over them, the 1990s have seen new strategies which find the contexts indicative of each sense in a training corpus, and then find the best match between those contexts and the instance of a word to be disambiguated.

Evaluation

As a result, there are now quite a few working WSD programs. One obvious question is therefore which is the best? Evaluation has excited a great deal of interest across the language engineering world of late. Not only do we want to know which programs perform best, but also the developers of a program want to know when modifications improve performance, and how much, and what combinations of modifications are optimal. US experience with ARPA's competitive evaluations for speech recognition, information retrieval, etc. has been that the focus provided by an evaluation serves to bring research communities together, forces consensus on what is critical about the field, and leads to the development of common resources, all of which then stimulates further rapid progress.

Reaping these benefits involves overcoming two major hurdles. The first is agreeing an explicit and detailed definition of the task. The second is producing a "gold standard" corpus of correct answers, so it is possible to say how much of the time a program gets it right. In relation to WSD, defining the task includes identifying the set of senses between which a programme is to disambiguate the "sense inventory" problem. Producing a gold standard corpus is both expensive (since it requires many person months of annotator effort) and hard (because evidence to date shows that different individuals or the same individual at different times will often assign different senses to the same word in context).

There is one gold-standard corpus in existence: the SEMCOR corpus. This comprises 250,000 words of text in which all content words have been tagged (manually) with the word sense. The sense inventory is taken from the WordNet lexical database. It is a very valuable resource and has already been widely used for WSD evaluation, but nevertheless it has several shortcomings. It is not big enough (there are only 83 words for which there are more than 100 sense-tagged corpus instances); WordNet, like any other dictionary, has various shortcomings, and these often result in anomalies in SEMCOR; and WordNet senses are rather finer-grained for many NLP tasks. There are also grounds for concern regarding the level of inter-annotator agreement.

The Resnik and Yarowsky proposals

Against this background, a workshop of the ACL Lexicon Special Interest Group (SIGLEX) in Washington earlier this year included a session on WSD evaluation. Philip Resnik and David Yarowsky presented their ideas, which sparked off a lively and productive discussion. The main sense of the meeting was that yes, there were great differences in people's theoretical perspectives, but there was also a job to be done, a job from which we would all benefit, and everyone present was willing to make compromises for the sake of a shared community view on how evaluation should proceed.

Resnik and Yarowsky put forward a framework in which, each year, a fresh subset of a huge corpus is used; one part of this is reserved for hand-tagging for evaluation, and the remainder released for training. A sample of, say, 200 ambiguous words (types, not tokens) is then chosen for use in evaluation. Each instance of each of those words in the evaluation subcorpus is manually tagged, thereby creating a gold-standard subcorpus in which just the instances of the sample words are tagged.

The community does not discover what the words are until their software is frozen for evaluation, so there is no risk of the software being optimised for those particular words. A new sample of test words is selected each year, and then last year's gold-standard corpus can be used for training and for improving programs.

Discussion

It soon became apparent that there were two cultures represented in the discussion: the computer scientists, who view a set of dictionary definitions as data they are to work with (and would like to be able to treat them as fixed) and the humanists, who have detailed experience of lexicography and textual analysis, and whose dominant concern lay in the sheer difficulty of identifying and defining word senses.

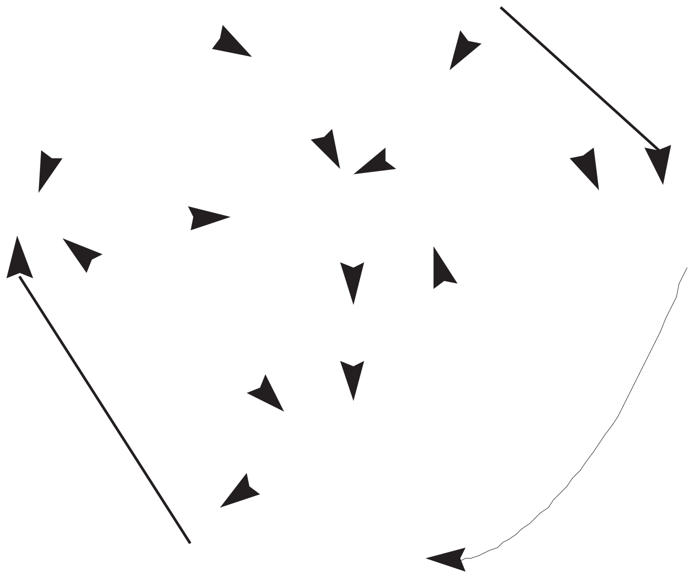
The humanists argued that high inter-annotator agreement was hard to get because existing dictionaries were not up to the task. This is scarcely surprising: they were written, for the most part, to explain word meanings to people, not to make cut-and-dried distinctions between senses. Nevertheless, without high inter-annotator agreement, the gold standard was fool's gold. There would only be potential for such agreement if the dictionary and its sense inventory were of a very high quality, and designed for the purpose. This could be achieved through allowing the people who were doing the tagging to improve the dictionary entry, perhaps changing the senses for the word, if they found that the corpus data they were tagging was at odds with the input dictionary (at least from an NLP perspective). They could also make much fuller dictionary entries as they would not be constrained to column inches, as paper lexicographers always are. In the Resnik-Yarowsky proposals, just 200 test words would be worked on each year, which suggested a manageable amount of lexicography revision to undertake year on year.

Allowing shifting goalposts in the form of a revisable sense inventory makes for great difficulties for WSD algorithms. But to be endorsed by the research community, an evaluation framework must not only provide computable measures, but must be valid. For this, a fully defensible sense inventory and gold standard are essential.

Other resolutions taken in the course of the meeting were that WordNet would be the starting point for the sense inventory; that the part-of-speech tagging task should be separated from the WSD task; that a positive score should be assigned to near misses and to ambiguity reduction (rather than all taggings simply scoring 1 or 0); and that the debate should continue over the SIGLEX e-mail list, with a view to beginning the annual cycle as outlined by Resnik and Yarowsky in the not-too-distant future.

For a fuller version of this paper (including references), see:
http://www.itri.bton.ac.uk/Adam.Kilgarriff/wsd-eval.ps.gz

Figure 1: Framework for gold-standard sense-tagged corpus generation and WSD evaluation



Adam Kilgarriff
Senior Research Fellow
Information Technology Research Institute
University of Brighton
Lewes Road
Brighton BN2 4GJ
United Kingdom
Phone: (44) 1273 642900 (ext. 919)
Fax.: (44) 1273 642908
E-mail:Adam.Kilgarriff@itri.bton.ac.uk
http://www.itri.bton.ac.uk/Adam.Kilgarriff

The ELRA Marketing Survey

The following article gives a brief overview of ELRA's planned marketing survey, with presentations of the objectives and the approach. The finalised version of the full text will be sent to all ELRA members no later than the beginning of September, when you will be asked for comments and input.

As part of its remit to promote the collection, dissemination and marketing of language resources, ELRA is preparing to conduct a market study in the language engineering/language resources field. When completed, the study will allow us to better understand the situation of the market in Europe and users' expectations. In addition, the facts gathered can be used to set up detailed performance targets for ELRA, including revenue/cost estimates and time scales for future activities. What is more, the entire language engineering field will be able to benefit from the analysis performed in the study. For one thing, the survey will provide information on users' real needs which can be passed on to the producers of the data. For another, where the need for particular language resources or data which are expensive to produce is established, ELRA can contact a number of funding bodies in order to help initiating the production.

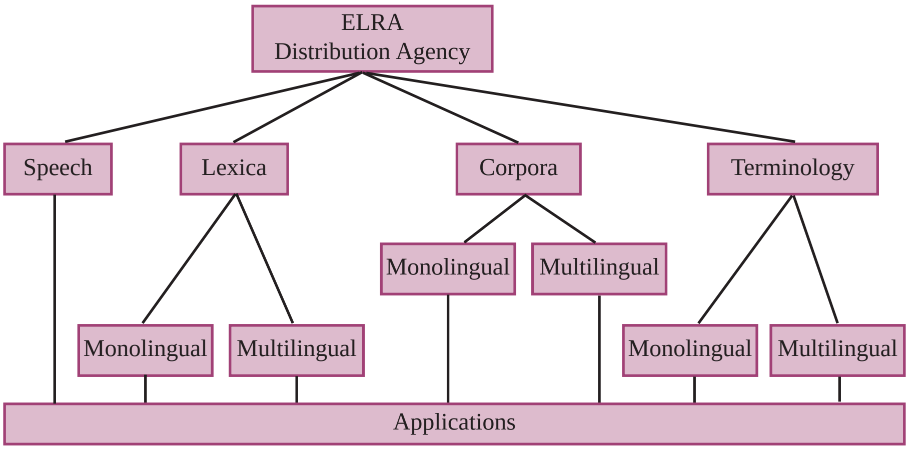
The primary objectives for the ELRA market study are as follows.

- First, the aim is to define/identify the current and future market structure - i.e. we want to obtain accurate figures for each of the market segments in which we are involved, and on the composition of each segment (profile of key players, key applications, trends). Other key figures required are the size of the market per segment, per geographical area, and world-wide.
- Second, we want to obtain a clear picture of users' needs and expectations, in order to be able to plan future activities and developments.
- Last but not least, we want to end up with an overview of pricing conventions and market rules from the various market players.

Our approach to gathering information is to start by establishing the background. To do this, we shall present structures of how we look at the market, including how we see ELRA's role on it, the resources and applications which already exist today, and how ELRA should approach the various different players. In a second stage, a field study will be performed using a number of different techniques, including interviews, questionnaires, analysis of corporate policies, publication analysis, etc. All mechanisms will be employed in order to reach as many market players as possible, and to ensure that the information collected is as accurate as it can be.
Figure 1 shows our view of ELRA's market role as a distributor, and the resources which ELRA covers (Speech, Lexica, Corpora and Terminology). The last three have been subdivided for practical purposes into monolingual and multilingual areas.

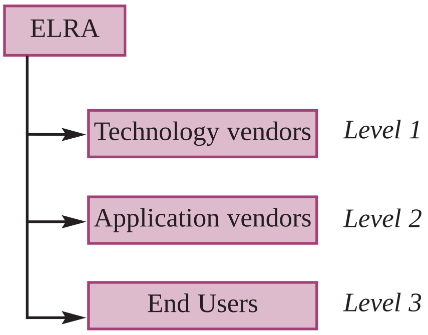
The applications belonging to each of the different resource areas are listed and defined more specifically in the complete version of the study, and comprise either applications for end users or tools (e.g. software blocks) which can be integrated with other tools to produce a complete package/application. For the time being, the different types of applications are mixed up, since the objective is to list every possible application which can be based on each resource.

Figure 1: ELRA's market role as a distributor



As can be seen in Figure 2, ELRA has divided the target market into three levels - technology vendors, application vendors and end users - on the basis of the different users involved. The main targets as far as ELRA is concerned are to be found in levels 1 and 2, as the number of actors on the end-user level makes it too difficult to reach them successfully with the resources at our disposal. However, the players included in levels 1 and 2 can be considered to function as a link between ELRA and end users.

Figure 2: Target market



We have now described in some detail how we have tried to segment the market, but we would ask you to comment on the structure we have developed and the way in which the market study has been set up.
More specifically, when you receive the full text, we would like to know if you think that we have exhausted all possible applications for each resource. Alternatively, have some applications been incorrectly assigned according to your view of the market?
We would also like comments on the key players/market leaders involved in the different applications. Last but not least, we would like your comments on our intention to mainly target Levels 1 and 2 of the market.
If you may have any ideas or comments on the structure of the study so far, we would appreciate hearing from you.
Thank you very much in advance.

For any comments or for more information, please contact:
Malin Nilsson
ELRA/ELDA office
Phone: +33 1 45 86 53 00
Fax.: +33 1 45 86 44 88
E-mail: elra@calva.net

New resources

ELRA-S0038 Siemens VoiceMail (American English)

VoiceMail consists of 17.5 hours of read speech (divided into 9.5 hours of transliterated speech and 8 hours of non-transliterated speech), recorded over the digital telephone network (ISDN) with 921 speakers originated from the USA (mainly native speakers and over 18 years old). It contains orthographic transliteration for about 25,000 utterances (out of 34,912 utterances in total). It has been designed in particular for telephone applications.

Language: American English File format: 16 bit linear Sampling rate: 8 kHz Speakers: 377 male and 544 female
Medium: 2 CD-ROM Size: 9.5 hours of transliterated speech, 8 hours of non-transliterated speech
Standard in use: headerless, one separate transliteration file comprising all utterances of all speakers

ELRA-S0039 APASCI (ITC-IRST) is an Italian speech database recorded in an insulated room with a Sennheiser MKH 416 T microphone. It includes 5,290 phonetically rich sentences and 10,800 isolated digits, for a total of 58,924 word occurrences (2,191 different words) and 641 minutes of speech. The speech material was read by 100 Italian speakers (50 male and 50 female). Each of them uttered 1 calibration sentence, 4 sentences with a wide phonetic coverage, 15 or 20 sentences with a wide diphonic coverage. Six of these speakers (3 male and 3 female) read 26 occurrences of the calibration sentence, 104 sentences with a wide phonetic coverage, 390 sentences with a wide diphonic coverage. 54 of the speakers (42 male and 12 female) pronounced 20 repetitions of the 10 isolated digits. The documentation of the database includes the transcription of each sentence both at phonemic and at orthographic levels.

This database allows to design, train and evaluate continuous speech recognition systems (speaker independent, speaker adaptive, speaker dependent, multispeakers). It was also designed for research on acoustic modelling as well as on acoustic parameters for speech recognition and for research on speaker recognition.

Format: 16 bit linear Standard: NIST SPHERE Sampling rate: 16 kHz Medium: CD-ROM

ELRA-S0040/ELRA-S0041 Danish SpeechDat(M) database - The Danish SpeechDat(M) database is the speech database collected within the SpeechDat(M) project. It consists of polyphone-like data recorded by 1,523 speakers. The speech files are stored as sequences of 8 bit 8 kHz A-law samples. Each prompted utterance is stored within a separate file and the associated label files are stored in SAM file format. An ASCII file is attached and is listing information about each speaker: speaker code, sex, age, region, prompt number. The lexicon is presented in a TAB delimited ASCII file containing an alphabetically ordered list of distinct lexical items occurring in the database. Each entry contains a frequency count and corresponding pronunciation information.

Example:	WORD	FREQUENCY	PHONEMIC TRANSCRIPTIONS
	åbnede	104	O b n @ D | O b n @ D @
	adresseangivelse	97	a d R a s @ a n g i: u l s @

The complete Danish SpeechDat database consists of 5 CD-ROMs. The first three CD-ROMs contain the application oriented sub-set.. The last two CD-ROMs contain the phonetically rich sentences.

The included items are: 5 application word phrases (semi spontaneous), 12 connected digit strings with 8 digits, 24 natural numbers (3-4 digits), 27 application words, 3 dates with D3 spontaneous (birthday), 3 spelled words, 2 money amounts with M1 small and M2 large City name (spontaneous), 3 yes/no questions (spontaneous), 22-25 sentences, T1 time phrase, T2 time of day (spontaneous).

There are 1,523 speakers in the SpeechDat database from 11 linguistic regions of Denmark and five age groups (under 16, 16-30, 31-45, 46-60, over 60). 78% of them are between 16 and 60 years old.