# The ELRA Newsletter

January - March 2000

Vol.5 n.1

## Contents

***Letter from the President and the CEO*** _____ *Page 2*

***ELRA Annual Report 1999***
*Khalid Choukri* _____ *Page 3*

***The GEMA Project***
***Gate for an Enhanced Multilingual resource Access*** _____ *Page 4*

***EC Call for Evaluators and Reviewers*** _____ *Page 4*

***LREC 2000 News*** _____ *Page 5*

***Applied Speech Processing Technologies - Our Journey***
*Siegfried Kunzmann* _____ *Page 6*

***EuroWordNet: a Multilingual Database with Wordnets in 8 Languages***
*Piek Vossen* _____ *Page 9*

***Example-Based Machine Translation at Carnegie Mellon University***
*Ralf Brown* _____ *Page 10*

***New Resources*** _____ *Page 13*

*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:***
Jeffrey Allen

***Layout:***
Audrey Mance

***Contributors:***
Ralf Brown
Khalid Choukri
Siegfried Kunzmann
Piek Vossen

*ISSN: 1026-8200*

***ELRA/ELDA***
*CEO: Khalid Choukri*
*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*
*WWW: http://www.elda.fr*

# *Dear Members,*

The Members' Annual General Assembly is one of the main events for ELRA during the first quarter of the year 2000. It will take place on 27 March at the "Grande Arche de la Défense" in Paris. We provide in this issue of the ELRA newsletter a short report on ELRA's activities in 1999 as well as objectives set for 2000 and beyond. By now, our members should have received detailed reports regarding ELRA technical activities and financial aspects.

A new call for proposals was launched by ELRA on 15 January 2000 for the production and packaging of modern French corpora. This call fits within the framework of on-going collaboration between ELRA, its distribution agency ELDA, the French Ministry of Culture ("Délégation générale à la langue française"). Such efforts to increase our collection of French language corpora will allow for improved language processing for this language in the future.

The kick-off meeting for the GEMA project took place on 10 January 2000. This project aims at setting up a web portal for terminology, language resources and all services related to Language Engineering. A survey has been created and is now available at http://www.elda.fr/proj/gemasurv.html in order to better determine user needs for this portal.

The Language Resources and Evaluation (LREC2000) program is being updated regularly. The conference will take place from 31 May to 2 June 2000 in Athens, Greece. Nearly 300 papers and posters have been accepted; the full list of presentations is currently available at http://www.elda.fr/lrec2000.html under the section entitled "Provisional Program". Information on the 11 satellite pre- and post-conference workshops is also available at the Web site under the section "Satellite Workshops". Many R&D systems for Natural Language and Speech Processing will be presented at the LREC2000 Exhibition which will take place in parallel with the conference. Please note that exhibit booths are available for those wishing to present their systems. More information on the LREC2000 Exhibition can be obtained from either Khalid Choukri (choukri@elda.fr) or Stelios Piperidis (spip@ilsp.gr).

In this issue of the Newsletter, we include the report on ELRA 1999 activities, the European Commission call for evaluators (mentioned above), and several articles relating to Language Resources. The first article by Siegfried Kunzman of IBM Speech Systems, traces the history of speech technologies and gives some indications with regard to their future applications for IST. The second article, written by Piek Vossen of Sail Labs, provides an update on the EuroWordNet resources that are available via ELDA. The final article in this issue, by Ralf Brown of the Language Technologies Institute of Carnegie Mellon University, presents the Example-Based Machine Translation approach to working with language resources.

As usual, the final section includes a list of newly acquired Language Resources:
- ELRA-S0076 French SpeechDat(II) FDB-5000;
- ELRA-S0077 Telephone Speech Data Collection for Czech;
- ELRA-S0078 Finnish SpeechDat(II) FDB-1000;
- ELRA-S0079 Finnish SpeechDat(II) FDB-4000;
- ELRA-S0080 Finnish-Swedish SpeechDat(II) FDB-1000;
- ELRA-W0021 ICE-GB (British English component of the International Corpus of English);
- ELRA-W0022 ILSP/ELEFTHEROTYPIA Corpus (PAROLE Greek Corpus);
- ELRA-L0032 PAROLE Greek Lexicon.

Prices are also announced for ELRA-W0020 Corpus French PAROLE. Please also note the new prices for LantMark lexica included in this issue.

We would like to remind you that some sections of the ELRA Web site, as well as the monthly Members' News bulletin are designated specifically for ELRA members. Please check and see if your institution has renewed its membership for 2000. If not, your institution is unfortunately no longer entitled to ELRA member benefits. Further information pertaining new memberships and membership renewal can be obtained from the ELDA office by contacting choukri@elda.fr or mapelli@elda.fr

Last but not least, in order to evaluate upcoming project proposals and to review current projects, the European Commission recently announced a call for tenders for recruiting evaluators for IST and post-MLIS projects. The full announcement of the call, as well as additional information, are available on page 4 of this ELRA Newsletter issue. We encourage our readers to visit the Internet site http://www.linglink.lu/htl for all information concerning the HLT calls.

Antonio Zampolli, President

Khalid Choukri, CEO

# ELRA Annual Report 1999

*Khalid Choukri, ELRA CEO*

Our activities continue to expand in different directions and areas. Our distribution efforts are quite stable in revenue despite a smaller number of resources sold (we managed to distribute 168 resources in 1999 compared to 210 items in 1998). Our catalogue of resources shows over 558 resources compared to 535 one year earlier. Our member base consists of 74 paid-up members in 1999 compared to 72 in 1998 and 66 in 1997. We managed to regularly publish our newsletter, one issue per quarter as planned. The 1999 call for proposals for Language Resources Packaging and Production was very successful and attracted about 30 proposals out of which 8 are being partially funded (resources available by March 2000). ELRA conducted a number of surveys which will help us better plan our long term activities. Our members benefited from such surveys through the reports we made available.

ELDA has started a new service on data collection which targets all users of LRs, in an attempt to help them produce or outsource to ELDA the production of the resources they need. The validation activity proceeds in a satisfactory way: we implemented the validation of a set of corpora and lexica produced within the Parole project and we started establishing our Network of Validation Units, through a public call for application in the Speech area as a first step.

ELDA has submitted two proposals within the MLIS programme (GEMA, Network-DC) which were accepted and started in January 2000, in addition to active involvement in other FP5 projects. ELDA has and will continue to benefit from grants and projects with the French government.

LREC-2000 organization is in an advanced stage and we expect to have a very successful conference in Athens next May/June 2000 (see details page 5).

ELDA/ELRA have also contributed to evaluation projects in the Information retrieval area and for speech recognition technologies. We should be able to launch a new sector of activities on LE and HLT evaluation for the benefit of the whole Language Engineering Community.

## Distribution of Language Resources

During this fiscal period we kept our sales at the level of 1998. The number of resources sold in 1999 is 168, compared to 210 in 1998. Despite the decrease in number of items, our revenues have grown. During the last fiscal year 98-99 (15 months) our sales in the Speech area represented over 85.9%, while written area (corpus/lexicon) represented about 14.1%, which is very similar to the figures of the previous fiscal year (1997-1998 12 months), respectively 86.6% and 13.4%. Our contribution to R&D efforts is also very stable both in percentage of revenues and in the number of items distributed to R&D labs: about 6.7% of revenues for 100 items (over the 15 months of the fiscal period) or 6.41% for 86 items over the 12 months of 1999. Sales for commercial purposes represents 93.3% for 94 items (over 15 months) or 93.6% for 82 items (over 12 months).

## Identification of Language Resources

As our core business is identification of new resources, we devoted a substantial effort to finalize a number of new agreements, in particular with EuroWordNet producers (EuroWordNet proved to be one of our top-selling resources). We also decided to remove a number of resources from the catalogue so we had to terminate a number of agreements (e.g. supply of low quality data compared to the samples initially received, acquisition of providers by other companies with a different strategy, etc.). The total of resources removed was 21 speech databases, 20 multilingual lexica, and 89 terminology databases. The catalogue issued in September 1998 consisted of 105 speech resources, 189 written resources (both lexica and corpora), and 361 terminology databases. During this fiscal period, we also succeeded to secure over 17 speech resources, 10 written resources, and 3 terminology databases which led to a catalogue of 102 speech resources, 20 written corpora, 48 monolingual lexicons, 113 multilingual lexicons and about 275 terminology databases by 31st December 1999.

## Validation of Language Resources

Our pilot application regarding the validation of a sub-set of resources produced within the Parole project proved to be very useful. Our goal was to validate a few resources (basically Spanish, Italian and Danish) and to assess the applicability of our manuals and procedures. The validation has been completed for Spanish and Italian; Danish will be completed in the first quarter of 2000. We will have to revise our manuals during year 2000 according to the feedback received from the Validation centers. We have succeeded in improving the quality of resources that we expect to constitute some of our best sellers.

During this period, we started establishing our network of validation units, and SPEX (Centre for Speech Processing Expertise, the Netherlands) has been selected after an open call, to act as an ELRA Validation Center for Spoken Language Resources.

## Commissioning the Production of Language Resources

Following the ELRA call issued in 8 February 1999, we received 29 proposals. All have been reviewed by three independent experts. The board of ELRA and a representative of the European Commission have acted as the selection committee on the basis of the experts written reports. 8 proposals have been selected for partial funding from ELRA, for a budget of about 200 K€, allocated from the LRs-P&P budget. We expect to get these resources by March 2000.

## ELRA Membership

Throughout the last three years we have noticed a global steady membership base. If we consider the sectors of activities (speech, text, terminology), we notice a particular decrease in the terminology sector and a relatively important increase in the speech sector. This year we had about 95 members (including those who did not pay yet their membership fee). The paid up members are 44 members in speech, 22 in written and 5 in terminology, compared to, respectively, 40, 24, 8 for 1998. We can also raise that out of the 95 members of 1999, 23 joined ELRA since 1st January 1999.

## Promotion and Awareness

ELRA continues to promote its activities at the major conferences and fairs. We started the preparation of LREC-2000 which looks very promising. We also continue to issue the ELRA newsletter four times a year with pages devoted to describe our new resources. The ELRA newsletter is published in French and English. Our web site

has seen an impressive number of visitors, in particular for the catalogue pages.

### Relationship with the European Commission

The first European project that helped ELRA establish its infrastructure ended in September 98. A new project (LRs-P&P - LE4-8335), awarded to ELDA in the framework of the last call of FP4, due to start in June 98, has been shifted to start in September 98 to avoid any perception of possible duplication of work. The project aims at monitoring the Language Engineering market (in particular Language Resources aspect) and commissioning the production of some key resources. The ELRA'99 call for Language Resources Packaging and Production has been initiated in the scope of this project.

### Future Work

ELRA/ELDA will carry on their regular activities related to the identification of new resources, the distribution, and the sales. We will continue to promote our activities through the quarterly newsletter and other information dissemination means. A specific and targeted marketing action following the users analysis and market monitoring (as a follow up of LRs-P&P) will be conducted to update our business and investment plans. We also need to liaise with the new IST projects that plan to produce Language Resources, to reach an agreement on distribution issues. ELDA will be actively involved in the GEMA and NETWORK-DC projects (MLIS program) and we will make sure that these projects are managed with high quality standards. A number of projects submitted to the French government were accepted for funding and we need to put more efforts on these projects for which we will need to recruit new employees.

ELRA has started contributing to evaluation programs through the supply of Language Resources, appropriate for evaluation and testing. ELRA has also been involved in the ELSE project as a non-funded partner. It is important to envisage that ELRA extends its activities towards evaluation and officially starts a new branch of activities related to Evaluation. This should also apply to Multimodal and Multimedia resources.

We will also continue the implementation of our Language Resource Validation work, in particular the work already planned by our validation Unit (SPEX).

The organization of LREC-2000 will also constitute a substantial effort on which ELDA staff will have to focus. It is important that we continue to organize a very high quality event both in terms of technical content and organizational issues.

Khalid Choukri
ELRA / ELDA
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

## GEMA - Gate for an Enhanced Multilingual resource Access

The aim of the GEMA project (Gate for an Enhanced Multilingual resource Access) is to provide a central and organised access point for the linguistic sector and to build and develop a linguistic portal with the corresponding services. Those services will cover a large range of activities, disciplines and needs of this sector and will include: on-line resource consultation services, on-line resource and tool acquisition services, information services, forum services and value-added services. GEMA has been conceived with the latest technologies in terms of Web developments and relies on the strong experience of some of its partners in the language sector. From its very first design, the project will clearly focus of the users' needs and will constantly search for their validation and feedback on the developments and functionality of the services.

The first step in this project consists of studying and specifying the needs expressed by the different types of users of the portal. Following this analysis, the specifications of all the developments will be carried out, from the functionality, services to be developed to the final exploitation plans.

**If you have still not filled out the Users' Survey, please visit:**
**http://www.elda.fr/proj/gemasurv.html**

## EC Call for Evaluators and Reviewers

The EC has issued a call for evaluators and reviewers "since a proper evaluation of upcoming IST and post-MLIS calls depend on the availability of a better, broader skills base". The EC has specifically mentioned that it needs evaluators with experience in "(a) industrial research and product development, (b) near-market, applications-oriented RTD and take-up actions, (c) management and business oriented aspects of projects, and (d) industrial representatives for lesser countries."

More information pertinent to this call is available at: http://www.linglink.lu/hlt Those interested in serving as evaluators/reviewers and who have NOT yet registered with the EC Cordis web site (http://www.cordis.lu), should do so at their earliest convenience. Any queries relating to this should be addressed to: *<evalexperts@dg12.cec.be>*

(1) PDF application forms can be downloaded at:
*http://www.cordis.lu/expert-candidature/home.html*

(2) The online registration facility can be accessed at:
*http://candidature.cordis.lu/expert-evaluators/*

**For more information, please contact Roberto Cencioni at**
INFSO-D4
European Commission, Information Society DGXIII
Euroforum 0-176, Jean Monnet Building
Rue Alcide de Gasperi
L-2920 Luxembourg
Tel.: +352 4301 32886
Fax.: +352 4301 34999
E-mail: Roberto.Cencioni@cec.eu.int

# LREC 2000 News

## Language Resources and Evaluation Conference
### Conference Date: 31 MAY - 2 JUNE 2000

About 300 oral and poster prsentations have been accepted by the LREC-2000 Programme Committee. They are listed according to four main domains (S: Spoken resources and Evaluation areas, W: Written area, T: Terminology area, E: Evaluation within Written area) at the following address: http://www.elda.fr/lrec2000.html.

The structure of the workshops is given below:

**o Pre-Conference Workshops**

| | 29 May 2000 | | | |
|---|---|---|---|---|
| 8:00 | **Workshop 1** | **Workshop 2** | | |
| | L. Dybkjaer, "From spoken dialogue to full natural interactive dialogue. Theory , empirical analysis and evaluation" | C. Draxler, "Very Large Telephone Speech Databases", 1st part | | |
| | *Please note that times and breaks still to be determined* | | | |
| 13:30 | **Break / Lunch** | | | |
| 14:30 | **Workshop 3** | **Workshop 2 Con't** | **Workshop 4** | **Workshop 5** |
| | P. Wittenburg, "Meta-descriptions and annotation schemas for multimodal/multimedia language resources", 1st part | C. Draxler, "Very Large Telephone Speech Databases", 2d part | K-S. Choi, C. Galinski, "Terminology resources and computation" | B. Maegaard "Workshop on the Evaluation of Machine Translation" |
| 20:00 | | | | |

| | 30 May 2000 | | | |
|---|---|---|---|---|
| 8:00 | **Workshop 3 con't** | **Workshop 6** | **Workshop 7** | **Workshop 11** |
| | P. Wittenburg, "Meta-descriptions and annotation schemas for multimodal/multimedia language resources", 2d part | J.McNaught, "Information extraction meets corpus linguistics" | E. Efthimiou, "Language resources in educational applications" | J. Mariani (CLASS project), "Using Evaluation within HLT Programs : Results and Trends" |
| | *Please note that times and breaks still to be determined* | | | |
| 13:30 | **Break/Lunch** | | | |
| 14:30 | **Workshop 8** | **Workshop 9** | | |
| | N. Ide, " Data Architectures and Software Support for Large Corpora: Towards an American National Corpus" | B. Williams, "Developing LR for minority languages: re-usability and strategic priorities" | | |
| 20:00 | | | | |

**o Post-Conference Workshops**

| | 3 June 2000 |
|---|---|
| 8:00 | **Workshop 10** |
| | J. Tsujii, “The integration of domain specific knowledge sources in NLP applications” |
| | *Please note that times and breaks still to be determined* |
| 13:30 | **Break / Lunch** |

**o Panels:** A set of panels are going to be organised during the Conference. One of the panels, organised by A. Zampolli, is entitled “Funding Agency and International Cooperation”; a second one, organised by Z. Vetulani, deals with “Human Language Technology Resources for Central European Languages: European Integration Issues”. More information will be made available on the LREC-2000 web site as soon as possible.

For information with regard to exhibiting at LREC2000, please contact the LREC2000 Conference Secretariat at:

**LREC2000@ilsp.gr**

The cost of stands for the whole Conference duration are available at:

**http://www.elda.fr/lrec2000.html**

For other technical information, please visit:

**http://www.elda.fr/lrec2000.html**

# Applied Speech Processing Technologies - Our Journey

*Siegfried Kunzmann, IBM Speech Systems*

## Introduction

Speech processing technologies have improved tremendously over the last decade and especially during the very last few years. This progress now allows the integration and use of spoken input to solve real world tasks on PCs, consumer devices (e.g. PDAs, mobile phones, appliances, automotive environment, etc.) as well as over (landline, wireless) telephone lines (e.g. directory assistance services, information retrieval tasks). Public attention to the progress of speech technologies has continuously grown, especially after the commercial introduction of highly accurate, general purpose, large vocabulary, speaker independent, continuous speech recognition systems for the PC (retail) market to a rapidly growing number of users in 1997 by Dragon and IBM, and soon followed by systems from L&H and Philips. In the general public this awareness of speech also raised lots of expectations regarding the possibilities of integrating accurate, large vocabulary speech recognition and understanding into different, complex application scenarios - by far not limited to applications allowing to produce text via voice. These growing expectations rapidly led to a change in focus of speech researchers from exploring speech technologies independently towards the building of complete (conversational) systems solving real world tasks by integrating and making available all required speech technologies (like recognition, understanding, dialog management, TTS). This, in turn, changed the common assessment of speech driven applications: from pure accuracy measurements towards improvements on supported functionality, user acceptance and usability from a human factors and natural interaction perspective.

Focusing on ViaVoice speech processing technologies we aim at giving an outline of the journey speech research has made so far and of the direction it is currently heading to facilitate and improve man-machine interaction. To illustrate the current state-of-the-art as well as future challenges we will describe technology needs within various application scenarios focusing first on progress in dictation systems and further on acceptance and usability improvements in telephony based man-machine interaction systems followed by the advantages in leveraging conversational technologies.

## Technology Progress for Dictation Systems

The first release of IBM's dictation system (IBM Speech Server Series™) in 1992 was targeted for a client-server environment to facilitate the production of text within large organizations or corporate environments. The speech processing services have been made available to each client via a TCP/IP based network from a workstation requiring special purpose hardware to support real-time recognition. Technology capabilities and computer resource limitations supported speaker dependent recognition of discrete speech with an active vocabulary of about 20,000 - 30,000 words. The user interface, called dictation window, provided basic functions like recording/playback, recognition, limited formatting and correction of the recognized text including voice-driven release of the dictated text to other applications. Soon after the initial releases technology progress allowed to make real-time speech recognition functions available on PCs supporting OS/2 or Windows (very recently also Linux and Macintosh). The first products on these platforms still required special purpose hardware to satisfy the intensive CPU demands of the speech algorithms and also to provide the functionality of a sound card. Rapidly increasing computer power, disk space, widely available (cheap) sound cards and microphones in conjunction with further progress on speech recognition technologies laid the grounds for making speech available to a largely growing user group. With the introduction of speaker-independent, continuous speech recognition in 1997, speech started to be widely accepted by users as a viable solution to increase productivity in day to day work.

User feedback has shown that accuracy related progress (e.g. further shortening enrollment phases, unsupervised acoustic adaptation, upfront analyzing of available user text, ongoing dynamic caching of dictated text, very large active vocabularies with > 200k words) is still a dominant factor for further research. Yet, it is at least as important, if not more, to address usability expectations and requirements from 'real' end users. For most of the users speech as alternative input device means seamless integration and support of widely used applications. End users, unlike early adopters, do not really want to be bothered with technology details. They just expect to 'get their text rapidly and painlessly into the computer'. During the last years user feedback led to significant changes in the dictation paradigm. Some recent examples for usability improvements are the introduction of features like direct dictation into standard applications, inline commands (e.g. casing, bold, italic), automatic formatting (e.g. categories like dates, numbers, currencies; spoken three dollar fifty, displayed 3,50$), voice driven navigation and correction, agents to help identify setup and recognition problems (e.g. microphone / audio setup, speaking style, how-to tutorials) and modeless dictation. Especially the introduction of modeless dictation improved the perceived user acceptance a lot. The user is now no longer required to say trigger words to switch between dictation of text or command execution. This achievement was possible by having the recognizer leverage short pauses ahead of commands and / or the context of the decoded utterance (command word sequences are within a grammar or free text) to automatically 'deduce' the intended (i.e. most likely) user action. Consequently, this resulted in a powerful feature to further improve the usability by consequently introducing a large number of natural, text manipulation commands like "print the next three paragraphs" or "select Ladies and gentlemen".

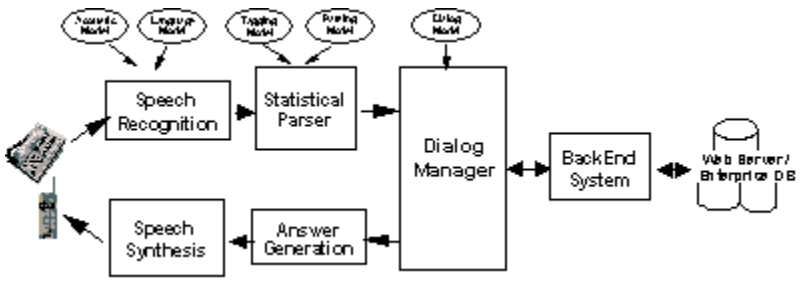


***Figure 1:*** *System Architectures*

Lowering acceptance thresholds and improving the usability will be the major tasks for the future in order to further establish speech as a viable, alternative input device for the desktop. Apart from obviously required improvements on basic recognition accuracy, the introduction of 'natural commands', implemented by extensive grammars, already points towards the desire of more flexible methods for generating text as well as for interacting with PC applications based on natural language understanding technologies.

## Technology Progress for Telephony Systems

For a rather long period of time highly optimized, accurate, small vocabulary (digit) recognizers have been available and deployed in large installations to allow speech-driven navigation of applications (e.g. directory assistance) designed for the phone pad, especially in telephony networks where touch-tone capabilities were not available. Typically, this resulted in a limited user acceptance, especially if the dialog structure gets more complex and complicated. The rapid progress in research on large vocabulary speech recognition technology and natural language understanding over the last few years made it feasible to apply these techniques also over telephone lines. Instead of saying (pressing) the number (button) 1 in a banking application to get information on the status of your account it is much more natural to assign words (e.g. account balance) or (grammar defined) word sequences (e.g. can you show me my account balance) that are meaningful for the user to trigger the required activity. In addition, the possibility to use different, dynamically generated grammars for different states of the application is a powerful first step to increase the usability and provides capabilities to handle complex applications like information retrieval tasks which require access to dynamically changing database contents. Systems with clearly defined menu structures we call directed dialog systems.

Directory Dialing, a directed dialog system available for various languages, is a typical representative for a directory assistance service application where people call a central number to be connected directly to an employee in a large corporation or to receive information about alternative ways of getting in touch with that person e.g. via e-mail, cell phone, fax. Directory sizes range from a few thousand names up to several hundred thousands of names. In ear1y 1999, we set up a directory dialer in IBM's call center for North America. It is based on ViaVoice technology and manages a 200,000 names directory. In order to uniquely identify a person's phone number (or in general: information), the system typically asks for location and name. Also in 1999, for technology demonstration purposes, we set up a Directory Dialer based on the German telephone system managing access to over 1 million names. For certain very common names it is required, apart from stating the home town of a particular person, to give some additional distinguishing information like the street name in order to uniquely identify the request.

A large number of directory assistance applications is required to work in a variety of languages especially in a multinational, multi-lingual environment like Europe. Often these services are made available via different phone numbers, explicitly prompting the user to choose the spoken language or use language identification technologies to select the proper recognizer. By merging the phone set of several languages into a common phone set a unified, single acoustic system can be built, explicit prompts to the user can be avoided, which in turn minimizes the time for the user required to gain requested information. With such a multi-lingual acoustic system the actual recognition vocabulary can be defined during setup and definition of the application by specifying grammars containing the allowed words / phrases in each language. In this scenario each valid path through the grammar can be based on a different language. To demonstrate the effectiveness of such an approach we set up a traffic jam information system which allows the user to query the current traffic situation on German motorways in either British English, French or German over a single phone line. On request the traffic information is directly loaded from an internet server and prompted to the caller via speech synthesis in the user's language. This also requires the transformation of the (German) traffic information into a proper answer template for that language.

The introduction of speech driven telephony applications will rapidly grow over the next few years. This is not only triggered by the growing mobility of people and the demand for timely information (e.g. events, news, city services, travel and reservations, weather forecast) but also because of the need for further cost reductions to perform these information services 24 hours a day, 7 days a week.

## Increased Usability with Conversational Systems

In order to solve different tasks in a complex application by using a large vocabulary, directed dialog system the user is still constrained to follow a predefined menu structure designed for this application. This requires to give particular information at particular points in time; this is true even if the list of alternatives to choose from is rather long. Moreover, user access to telephony applications is typically relatively short and doesn't allow long 'learning curves'. Therefore, conversational system techniques which support functions like providing and requesting refined information, inheritance of information

across tasks, and mixed-initiative dialogs yield to much higher user acceptance. To gain such flexibility, system architectures as shown in figure 1 are required. To be able to deal with partial and even ungrammatical sentences, typical for natural interactions, grammar based systems are not powerful enough. Thus, the speech recognition component is characterized by the use of domain specific, statistical language models to decode the speech input. The Statistical Parser transforms the recognized utterance into an application specific slot representation. The Dialog Manager is carrying the dialog status during the entire conversation with the user, deals with refined information, takes initiative in case of insufficient information till the actual request to the Backend System can be initiated. The Backend System gets the requested information from a web server or directly from an enterprise database. The Answer Generation module generates natural prompts to structure information, to provide only partial information in case of too much alternatives, and incorporates retrieved information dependent on the dialog status or requested missing information. The actual transformation of the textual prompt into an audio message is handled via Speech Synthesis.

Extending the framework of directory dialing, we developed conversational dialers for British English, French, German and Spanish in 1999. The active dictionary sizes range from a few hundred names up to about 8000 names, e.g. in the French system which is currently deployed for internal use for IBMers in Paris. The main objective for these technology prototypes was to rapidly build, for various languages, conversational systems to demonstrate the benefits of natural language capabilities as well as to proove that common application design, concepts, functions and backend access can be shared (and partially unified) across languages, thus helping minimize the time to build the application. The currently supported functions of these prototypes are querying the number, dialing and the ability to leave messages for the addressee.

Common base design and concepts are especially required to allow enhancements of such conversational systems towards the integration of multi-lingual speech recognition technologies, so that unified access to the name dialing functionality can be made available for a variety of languages.

**SAMPLE DIALOG**

***System:*** IBM Name Dialing

***User:*** *Hallo, good morning, may I have the number of Carsten, please*

***System:*** We have Carsten Günther and Carsten Siegler.

**System Initiative: Desambiguation of names**

***User:*** *Give me the first*

**Selection from a short list**

***System:*** The number of Carsten Günther is 4 5 7 8.

***User:*** *Dial him*

**Resolving of pronouns**

***System:*** Dialing Carsten Günther 4 5 7 8.

While directed dialog systems using very large vocabulary speech recognition systems have by now become state-of-the-art, significant experience has been gained on the capabilities of conversational systems and especially on the increased usability and acceptance by users. Similar to the early versions of desktop dictation systems where very narrow-domain systems were built and deployed (e.g. radiology systems), natural language technologies will initially focus on specific applications and limited tasks, too. The technology is rapidly becoming more powerful with the target to get higher acceptance by users trying to handle complex tasks. Moreover, conversational technology will not only be used for desktop and telephony based applications but will also be required for kiosk systems.

## Outlook

In today's information society, multilingual content is made available to a rapidly growing population using the internet. In parallel, the increased introduction of mobile phones allows communication at any time and any place which in turn raises the demand for instant access to information. A lot of recent public statements predict the convergence of the traditional internet and the telephone network. Digital phone lines and Wireless Application Protocol technology combined with special browsers (e.g. optimized for small displays) allow to make internet content available via new information channels by extending the HTML paradigm which is based on having common application and business logic separated from presentation logic. Similarly, VoIP allows PC users to communicate over the Internet with other users connected to the telephone network. Emerging standards (WML, XML, VXML aim at simplifying and unifying application development across all these different input / output channels. This is complemented by users expecting such applications and technologies to be functional and easy-to-use. This requires the provision of common interaction models to on-line content and services across all information channels. Speech processing technologies (including high quality Text-To-Speech, transcription systems, machine translation, speaker identification / verification, …) will play a very dominant role to be able to fulfill unified access to information, especially if this framework is extended towards multi-modal interfaces that require the integration of gesture recognition and speech processing. Apart from the technological challenges this will also mean that speech technologies for the desktop, telephone and consumer devices need to be combined (e.g. local and client/server processing) and even merged in order to provide a robust, seamless, speech driven user interface across all these input / output devices. This progress will help overcome language barriers by providing the ability to access all available types of on-line content and services with the same model of interaction.

Dr. Siegfried Kunzmann
Research & Technology Manager
European Speech Research, IBM Speech Systems
Vangerowstr. 18
D- Heidelberg
Tel.: +49 6221 59 4443
Fax: +49 6221 59 3500
E-mail: kunzmann@de.ibm.com

# EuroWordNet: a Multilingual Database with Wordnets in 8 Languages

*Piek Vossen, Sail-labs, Belgium*

EuroWordNet has been funded by the European Commission as projects LE2-4003 and LE4-8328 (http://www.hum.uva.nl/~ewn). The goal was to build a multilingual lexical database with wordnets for 8 European languages: English, German, French, Dutch, Spanish, Italian, Czech and Estonian. Each wordnet is structured along the same lines as the Princeton WordNet for English in terms of sets of synonyms, so-called synsets, between which basic semantic relations are expressed. For example, {car; auto; automobile; machine; motorcar} is a synset in WordNet that is related to:

- more general concepts or the hyperonym synset: {motor vehicle; automotive vehicle},
- more specific concepts or hyponym synsets: e.g. {cruiser; squad car; patrol car; police car; prowl car} and {cab; taxi; hack; taxicab},
- parts it is composed of: e.g. {bumper}; {car door}, {car mirror} and {car window}.

Each of these synsets is again related to other synsets thus constituting a huge network or wordnet. Such a wordnet can be used for making semantic inferences about the meanings of words (what meanings can be interpreted as vehicles), for finding alternative expressions or wordings, or for simply expanding words to sets of semantically related or close words in information retrieval.

The wordnets for each language in EuroWordNet are stored in a central lexical database system and each meaning is linked to a so-called Inter-Lingual-Index, thus creating a multilingual database. This index is based on the concepts in WordNet1.5, but has been adapted to provide a more efficient mapping. In the multilingual database it is possible to go from one meaning in a wordnet to a meaning in another wordnet, which is linked to the same index-record. In total, 90 different language-internal relations have been defined and 20 types of equivalence relations to the Inter-Lingual-Index.

Such a multilingual database is useful for cross-language information retrieval, for transfer of information from one resource to another or for simply comparing the different wordnets. Via the Inter-Lingual-Index, the wordnets also share a common top-ontology of basic semantic distinctions (such as (in)animate, natural, artifact). Figure 1 illustrates the modular multilingual design of the database. The industrial users in the project have validated the data and demonstrated its use in mono-lingual and cross-lingual information retrieval.

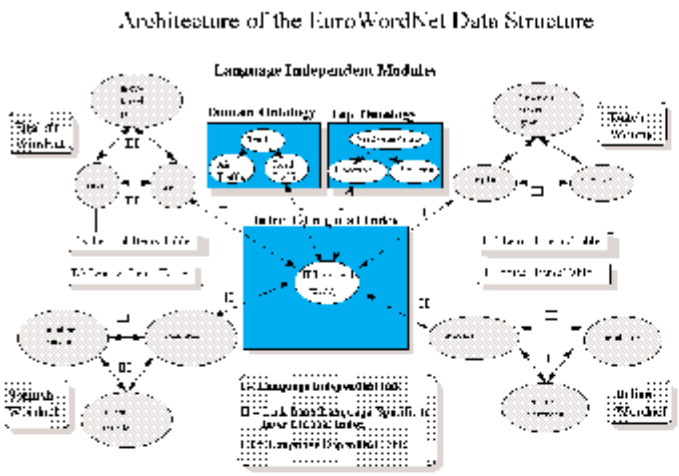


***Figure 1:*** *The multilingual design of the EuroWordNet database*

The data have been built using a common strategy starting from a shared set of 1300 most important concepts, the so-called Base Concepts. The Base Concepts (represented as records in the Inter-Lingual-Index) have been classified by a top-ontology of 63 semantic distinctions. This top-ontology provides a common framework for all the wordnets. Each language-specific wordnet contains a set of carefully selected mappings to these Base Concepts. The lexicalizations that are directly related to these Base Concepts (hyperonyms, hyponyms and other relations) have been specified manually. This resulted in core wordnets that have a high quality and are highly compatible across the languages. The core wordnets have been extended using semi-automatic techniques. Comparison of the wordnets has guided further improvements.

The EuroWordNet project finished in the summer of 1999 and the wordnets are available as plain text files and in database format. The database versions can be accessed, edited and compared in the multilingual database Polaris or viewed with the graphical interface Periscope. The wordnets and Periscope are distributed via ELDA/ELRA, the multilingual wordnet editor Polaris is distributed by Lernout and Hauspie (email: Geert.Adriaens@lhs.be). Both Polaris and Periscope run on Windows95/98/NT machines. The wordnets require between 10 and 25 MB disk space each. Another 70MB are needed for WordNet1.5 and the Inter-Lingual-Index. All data can however also be accessed from CD.

All project information and documentation can be downloaded from the EuroWordNet WWW-site, as well as free samples of the wordnets (as text files and as databases) and the Periscope viewer. The design of the database, definitions of the relations and structures are given in the general EuroWordNet document that can also be downloaded from http://www.hum.uva.nl/~ewn. The table page 10 gives a quantitative overview of the final wordnets.

We believe that EuroWordNet is a solid fondation for the development of language resources and technology that can be shared and transferred to all the associated languages. In addition to the use for (cross-language) information retrieval, there are many other applications that can

directly benefit from the multilingual semantic resources: information-acquisition tools, authoring-tools, language-learning tools, translation-tools, summarizers.

Finally, we expect that the EuroWordNet database will be extended to many more languages. Other groups are currently developing wordnets with national funds for other European languages using the same format and specifications as EuroWordNet. These wordnets can be linked to any other wordnet available in the database. Wordnets developed in collaboration with EuroWordNet cover the following languages: Basque, Catalan, Portuguese, Danish, Norwegian, Swedish, Romanian, Slovenian, Lithuanian, Russian, and Greek. An important aspect is here to maintain the framework so that the standardization effect will continue.

| | | Synsets | No. of senses | Sens./ syns. | Entries | Sens./ entry | LIRels. | LIRels/ syns | EQRels-ILI | EQRels /syn | Synsets without ILI |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Dutch Wordnet** | Nouns | 34455 | 54428 | 1,58 | 45972 | 1,18 | 84869 | 2,46 | 26724 | 0,78 | 6070 |
| | Verbs | 9040 | 14151 | 1,57 | 8826 | 1,60 | 25973 | 2,87 | 26724 | 2,96 | 1133 |
| | Other | 520 | 1622 | 3,12 | 1485 | 1,09 | 797 | 1,53 | n.a. | n.a. | n.a. |
| | Total | 44015 | 70201 | 1,59 | 56283 | 1,25 | 111639 | 2,54 | 53448 | 1,21 | 7203 |
| **Spanish Wordnet** | Nouns | 18577 | 41292 | 2,22 | 23216 | 1,78 | 40559 | 2,18 | 18634 | 1,00 | 0 |
| | Verbs | 2602 | 6795 | 2,61 | 2278 | 2,98 | 3749 | 1,44 | 2602 | 1,00 | 0 |
| | Other | 2191 | 2439 | 1,11 | 2439 | 1,00 | 10855 | 4,95 | n.a. | n.a. | n.a. |
| | Total | 23370 | 50526 | 2,16 | 27933 | 1,81 | 55163 | 2,36 | 21236 | 0,91 | 0 |
| **Italian Wordnet** | Nouns | 30169 | 34552 | 1,15 | 24903 | 1,39 | 83021 | 2,75 | 43848 | 1,45 | 98 |
| | Verbs | 8796 | 12473 | 1,42 | 6607 | 1,89 | 30757 | 3,50 | 27941 | 3,18 | 0 |
| | Other | 1463 | 1474 | 1,01 | 1468 | 1,00 | 3290 | 2,25 | n.a. | n.a. | n.a. |
| | Total | 40428 | 48499 | 1,20 | 32978 | 1,47 | 117068 | 2,90 | 71789 | 1,78 | 1561 |
| **French Wordnet** | Nouns | 17826 | 24499 | 1.37 | 14879 | 1.65 | 39172 | 2.20 | 17815 | 1.00 | 16 |
| | Verbs | 4919 | 8310 | 1.69 | 3898 | 2.13 | 10322 | 2.10 | 4915 | 1.00 | 4 |
| | Other | 0 | n.a. | n.a. | n.a. | n.a. | n.a. | n.a. | n.a. | n.a. | n.a. |
| | Total | 22745 | 32809 | 1.44 | 18777 | 1.75 | 49494 | 2.18 | 22730 | 1.00 | 20 |
| **German Wordnet** | Nouns | 9951 | 13656 | 1.37 | 12746 | 1.07 | 23856 | 2.40 | 10570 | 1.06 | 0 |
| | Verbs | 5166 | 6778 | 1.31 | 4333 | 1.56 | 10960 | 2.12 | 5762 | 1.12 | 0 |
| | Other | 15 | 19 | 1.27 | 19 | 1,00 | 2 | 0.13 | 15 | 1.00 | 0 |
| | Total | 15132 | 20453 | 1.35 | 17098 | 1.20 | 34818 | 2.30 | 16347 | 1.08 | 0 |
| **Czech Wordnet** | Nouns | 9727 | 13829 | 1.42 | 9277 | 1.49 | 19856 | 2.04 | 9729 | 1.00 | 0 |
| | Verbs | 3097 | 6120 | 1.98 | 3006 | 2.04 | 6403 | 2.07 | 3097 | 1.00 | 0 |
| | Other | 0 | n.a. | n.a. | n.a. | n.a. | n.a. | n.a. | n.a. | n.a. | n.a. |
| | Total | 12824 | 19949 | 1.56 | 12283 | 1.62 | 26259 | 2.05 | 12824 | 1.00 | 0 |
| **Estonian Wordnet** | Nouns | 5028 | 8226 | 1.64 | 7209 | 1.14 | 10873 | 2.16 | 5683 | 1.13 | 0 |
| | Verbs | 2650 | 5613 | 2.12 | 3752 | 1.50 | 5445 | 2.05 | 3321 | 1.25 | 0 |
| | Other | 0 | n.a. | n.a. | n.a. | n.a. | n.a. | n.a. | n.a. | n.a. | n.a. |
| | Total | 7678 | 13839 | 1.80 | 10961 | 1.26 | 16318 | 2.13 | 9004 | 1.17 | 0 |
| **English WordNet Addition** | Nouns | 4751 | 14188 | 2,99 | 2524 | 5,62 | 20707 | 4,36 | n.a. | n.a. | n.a. |
| | Verbs | 11363 | 25761 | 2,27 | 14726 | 1,75 | 21070 | 1,85 | n.a. | n.a. | n.a. |
| | Other | 247 | 639 | 2,59 | 70 | 9,13 | 363 | 1,47 | n.a. | n.a. | n.a. |
| | Total | 16361 | 40588 | 2,48 | 17320 | 2,34 | 42140 | 2,58 | n.a. | n.a. | n.a. |
| **WordNet1.5** | Nouns | 60521 | 107428 | 1,78 | 88175 | 1,22 | 159223 | 2,63 | n.a. | n.a. | n.a. |
| | Verbs | 11363 | 25768 | 2,27 | 14734 | 1,75 | 24331 | 2,14 | n.a. | n.a. | n.a. |
| | Other | 22631 | 54406 | 2,40 | 23708 | 2,29 | 27821 | 1,23 | n.a. | n.a. | n.a. |
| | Total | 94515 | 187602 | 1,98 | 126617 | 1,48 | 211375 | 2,24 | n.a. | n.a. | n.a. |

***Table:*** *Quantitative overview of the EuroWordNet database*

Piek Vossen
Sail-labs
Coveliersstraat 15
2600 Berchem, Belgium
Tel.: +32 3 287.64.50
Fax: +32 3 287.64.70
E-mail:
piek.vossen@sail-labs.be

**Explanation of the columns**

| | |
|---|---|
| Synsets | = concepts represented by synonymous word senses |
| No. of senses | = number of word senses, or synonyms |
| Sens./ syns. | = average of senses or synonyms per synset |
| Entries | = number of words |
| Sens./ entry | = number of senses per word |
| LIRels. | = number language-internal relations |
| LIRels/ syns | = average of language-internal relations per synset |
| EQRels-ILI | = number of equivalence relations to the Inter-Lingual-Index |
| EQRels/syn | = average of equivalence relations per synset |
| Synsets without ILI | = synsets without a equivalence relation |

# Example-Based Machine Translation at Carnegie Mellon University

*Ralf D. Brown, Carnegie Mellon University, Language Technologies Institute*

Example-Based Machine Translation (EBMT) is fundamentally translation by analogy. Given a corpus of pre-translated examples, one translates previously-unseen text by finding the best match(es) in the corpus and using the associated translation(s). If this sounds a lot like a translation memory (TM) system, it is because EBMT is a superset of translation memory.

There are a variety of approaches to finding the "best match"' for a new sentence. One can parse the corpus into syntactic parse trees and match the trees [5]. One can find the nearest single match for the complete input (as is typically done in translation memory systems) and then attempt to modify both the matched example and its associated translation to create an exact match for the input [6]. Or one can find the complete set of exact phrasal matches and piece together a full translation from the fragments, which is the approach taken at Carnegie Mellon for both the Pangloss and DIPLOMAT projects[2].

The advantage of using exact phrasal matching is that one does not need to build a parser (as would be the case for parse-tree matching) or implement language-specific modification rules. As a result, our EBMT system is perfectly suited for rapid development of new language pairs -- given a large sentence-aligned corpus such as the Hansard [4], an initial version of a new bidirectional translator can be created in one or

two days.

The disadvantage of exact matching is that it requires several million words of parallel text for reasonably broad coverage of unrestricted text, which may be difficult to obtain. Thus, we have embarked on a project (funded by the National Science Foundation) to add generalization of the pretranslated examples, thereby dramatically reducing the required size of the training corpus. These generalizations may be syntactic (noun, verb, adjective) or semantic (weekday, color, company, country, etc.), and are implemented by pattern replacement in both the corpus and the input to be translated.

The underlying idea behind generalization is that there are equivalence classes of words and phrases that may be used interchangeably in a particular context and still yield grammatical results. For instance, any pretranslated example which uses the word "Monday" could be modified to use "Tuesday" instead (or any other day of the week). At a more general level, if one can identify noun phrases in the source-language text, one can then (within limits) substitute any other noun phrase whereever a noun phrase occurs. One does not need expert linguists to create a full grammar of the language -- a small subset grammar capturing the most frequent patterns and phenomena will suffice; it can always be extended later as needs and available resources dictate.

The basic algorithm in the Carnegie Mellon EBMT system is to search the example base for the largest phrases from the input which are contained in each pre-translated sentence. For each match, the corresponding translation is determined by performing a word-level alignment [2] of the two halves of the translation example. The overall translation is assembled from the partial translations using a statistical language modeler in our multi-engine machine-translation architecture [1].

This partial exact matching is extended by allowing the equivalence classes mentioned above. Equivalence classes are applied by replacing any matching words or phrases with the name of the equivalence class, appending a disambiguating number if that equivalence class has already been used in the sentence (referred to as tokenizing in the remainder of this article). The process is repeated until no more replacements are possible, at which time a partial exact match against the example base is performed, just as previously without equivalence classes. Since the example base has also been tokenized, this allows interchangeable use of the members of an equivalence class.

In the input to be translated, phrases belonging to an equivalence class are always replaced by the class name. In the example base, the class members are only replaced if an appropriate translation is present in the target-language half of the example. To permit proper matching against the example base, ambiguous "words'" are permitted which match any of several alternatives at that location. Whenever a single word is replaced by its class name, the original word is retained as an alternative for matching; unfortunately, this is not possible for phrases as the difference in length would cause erroneous matches when examining the index. This capability for ambiguous terms also allows words to be in multiple equivalence classes provided that the translations are mutually distinct. (If there were a common translation between different equivalence classes, the system would be unable to decide which to use).

Whenever a term is replaced by its class name, the corresponding translation is remembered. Once a translation of the tokenized text has been found, each token is expanded by substituting the translation which was remembered when the text was initially tokenized. This back-substitution step yields the final translation which is output, and is what makes equivalence classes work.

As an example, consider the sentence

```
John Miller flew to Frankfurt on December 3rd.
```

This becomes

```
<firstname-m> <lastname> flew to <city> on <month> <ordinal>.
```

after an initial tokenization pass, and then

```
<person-m> flew to <city> on <date>.
```

after a second pass. The tokenized form will now match

```
Dr. Howard Johnson flew to Ithaca on 7 April 1997.
```

among many other possibilities.

A further generalization of equivalence classes involves repeated (recursive) matching against the example base. For this extension, translation pairs in the example base are tagged with a token which preferably contains linguistic information such as gender and number. Tagged entries are not limited to literal strings -- they may themselves contain tokens, allowing the use of paired production rules to create a grammar, as shown in Figure 1.

| | |
|---|---|
| ;;;(TOKEN <NOUN-M>) | ;;;(TOKEN <NP-M>) |
| book | the <NOUN-M> |
| livre | le <NOUN-M> |
| ;;;(TOKEN <NP-M>) | ;;;(TOKEN <NP-F>) |
| <POSS> <ADJ-M> <NOUN-M> | the <NOUN-F> |
| <POSS> <NOUN-M> <ADJ-M> | la <NOUN-F> |

***Figure 1:*** *Sample English-French Production Rules*

To perform a translation, the system first searches for phrases that completely match one or more tagged entries, and then substitutes the associated tags into the input text. This process is repeated until there are no more complete matches of tagged entries, at which point an extended form of the normal partial-exact match against all examples -- including tagged entries -- in the knowledge base is performed. As is the case when using equivalence classes, at each step the appropriate back-substitution is remembered so that it can be applied to the tokenized translation in order to produce the final output.

The process of matching against the corpus is more complex when grammar rules are involved, because not all alternative terms which will be matched represent the same number of words in the input. Each of the individual substitutions produced at any stage of the repeated tokenization described above may be matched against the corpus. Recursive matching permits a word to be in multiple equivalence classes even when the translations are not distinct, and can be applied more generally than simple tokenization because replacements are only made in the proper context.

When the tags contain linguistic information, this information can be used to enforce constraints and thus select the appropriate translation of a word. For the example shown in Figure 2, the English word "affordable" can be translated as either a singular or plural adjective; this is indicated by showing all alternatives for a given word as a list in parentheses. After a first recursive matching pass, both "affordable" and "painters" are tokenized. Searching the corpus for a further tagged match of this initial result yields only the masculine noun phrase in which both adjective and noun are plural, which disambiguates "affordable" as the plural form. Once no more matches are possible, the translation of the fully-tokenized

```
... the affordable painters ...
==>  ...  the  (<adj-s>  <adj-p>)
<noun-m-p> ...
==> ... the <adj-p> <noun-m-p> ...
==> ... <np-m> ...
          -=- translation into Spanish -=-
==> ... <np-m> ...
==> ... los <noun-m-p> <adj-p> ...
=> ... los pintores accesibles ...
```

***Figure 2:*** *Disambiguation through Linguistic Constraints*

input is determined, and the tokenization is reversed by back-substituting the appropriate translation for each tokenized term, as remembered during tokenization. The final result is a translation in which the correct alternative has been selected.

The effort of adding the grammar rules and linguistic information was quite modest, totalling an estimated 70-80 hours for the French system and 50-60 hours for the Spanish system. While the availability of morphological information for both French and Spanish considerably reduced the level of effort, for many language pairs much of the work can be performed automatically even without such data, given a bilingual dictionary which is required anyway. By matching suffixes or other lexical features, as was done for the Spanish system (and to a lesser extent for the French system), many of the most frequent morphological variations can be captured. Work is also underway on a method for automatically learning equivalence classes from the training corpus, which will significantly reduce the manual effort involved in adding linguistic information to generalize the corpus.

Adding equivalence classes produces a small but noticeable improvement, and the greater infusion of generalization due to recursive matches produces a greater improvement. A series of experiments published last summer [3] showed as much as an order of magnitude reduction in the amount of training text required to be able to translate a given percentage of arbitrary input. Our French system reaches 80% coverage of the test text with less than 300,000 words of training material (nearly all of which consists of grammar rules and morphological entries) when using recursive matching, but requires one million words without grammar rules and 1.2 million words when relying solely on the translation examples. Since the performance curve flattens out, the difference is even greater for higher coverage values -- achieving 90% coverage required less than half a million words with recursive matching versus 7 million words without. Similarly, the Spanish system reaches 80% coverage with only 350,000 words versus 2.5 million words and 90% with only 3 million words versus more than 11 million without generalization.

Translation quality is marginally lower when using the grammar rules, since it is easy to over-generalize. A further cause of reduced quality is that generalizations only produce a single, preferred translation, rather than a number of closely related translations that may vary according to context.

Figure 3 illustrates the effect of generalization on the system's output. For this example, the system was told to use very terse output: only the very best-scoring translation for any match is shown, and no matches which are entirely contained within another, larger match are shown. The left-hand column indicates which phrase was matched and the penalty score for the generated translation (zero is considered perfect), while the right-hand column shows the translation. With less than half as much parallel text (including morphological and grammar entries), the generalized version covers more of the input, with generally longer matches, but also with lower quality. Thus, “voix” is translated as “voice” rather than “votes” because the generalization rules do not take the nature of the corpus (parliamentary proceedings) into account. The much smaller match without generalization, on the other hand, correctly uses "votes" because that is the usage in the corpus.

Input:
La motion de M. Lewis est adoptée par 147 voix contre 77.
(Mr. Lewis' motion is adopted by 147 votes to 77.)

707,000-word corpus, no generalization:

| | |
|---|---|
| "la motion de M. lewis" (1) | "motion of Mr . Lewis" |
| "adoptée par" (0) | "adopted by" |
| "147 voix" (0) | "147 votes" |
| "77 ." (0) | "77 ." |

307,000-word corpus, with full generalization:

| | |
|---|---|
| "la motion de M." (2.21) | "the motion for Mister" |
| "motion de m . lewis" (0.4) | "Mister Lewis's motion" |
| "est adoptée par 147" (1.825) | "is adopted by 147" |
| "par 147 voix contre 77 ." (1) | "by 147 voice against 77." |

***Figure 3:*** *Comparison With and Without Generalization*

While increasing the effectiveness of the available translation examples is an interesting result for major languages such as French and Spanish, it is vital for languages which have little or no available parallel text; for such languages, being able to generalize the examples which can be found or manually generated may be what makes a translation system feasible at all.

## References


[1] Ralf Brown and Robert Frederking. *Applying Statistical English Language Modeling to Sym- bolic Machine Translation.* In Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-95), pages 221--239, Leuven, Belgium, July 1995. http://www.cs.cmu.edu/~ralf/papers.html.

[2] Ralf D. Brown. *Example-Based Machine Translation in the Pangloss System.* In Proceedings of the Sixteenth International Conference on Computational Linguistics, pages 169--174, Copenhagen, Denmark, 1996. http://www.cs.cmu.edu/~ralf/papers.html.

[3] Ralf D. Brown. *Adding Linguistic Knowledge to a Lexical Example-Based Translation System.* In Proceedings of the Eighth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-99), pages 22--32, Chester, England, August 1999. http://-www.cs.cmu.edu/~ralf/papers.html.

[4] Linguistic Data Consortium. *Hansard Corpus of Parallel English and French*. Linguistic Data Consortium, December 1997. http://www.ldc.upenn.edu/.

[5] M. Nagao. *A Framework of a Mechanical Translation between Japanese and English by Analogy Principle.* In A. Elithorn and R. Banerji (eds), editors, Artificial and Human Intelligence. NATO Publications, 1984.

[6] Tony Veale and Andy Way. Gaijin: *A Template-Driven Bootstrapping Approach to Example- Based Machine Translation.* In Proceedings of the NeMNLP'97, New Methods in Natural Language Processessing, Sofia, Bulgaria, September 1997. http://www.compapp.dcu.ie/~tonyv/papers/gaijin.html.



Ralf D. Brown
Carnegie Mellon University,
Language Technologies Institute
5000 Forbes Avenue
Pittsburgh, PA 15213-3890
E-mail: ralf+@cs.cmu.edu

# New Resources

## ELRA-W0021 ICE-GB (British English component of the International Corpus of English)

ICE-GB is the British component of the International Corpus of English (ICE). ICE began in 1990 with the primary aim of providing material for comparative studies of varieties of English throughout the world. Twenty centres around the world are preparing corpora of their own national or regional variety of English.

ICE-GB is fully grammatically analysed. Like all the ICE corpora, ICE-GB consists of a million words of spoken and written English and adheres to the common corpus design. 200 written and 300 spoken texts make up the million words. Every text is grammatically annotated, allowing complex and detailed searches across the whole corpus.

ICE-GB contains 83,394 parse trees, including 59,640 in the spoken part of the corpus.

ICE-GB has been fully checked. It was checked by linguists at several stages in its completion, using both a traditional 'post-checking' strategy and also by cross-sectional error-based searches.

ICE-GB is distributed with the retrieval software ICECUP (the International Corpus of English Corpus Utility Program). ICECUP supports a variety of query types, including the use of the parse analyses to construct Fuzzy Tree Fragments to search the corpus.

| **Price for ELRA members:** | | **Price for non members:** | |
|---|---|---|---|
| for research use: | 780 EURO | for research use: | 1,500 EURO |
| for commercial use: | 8,500 EURO | for commercial use: | 15,000 EURO |

## ELRA-S0076 French Speechdat(II) FDB-5000

The French SpeechDat(II) FDB-5000 comprises 5040 French speakers (2,693 females, 2,347 males) recorded over the French fixed telephone network. The SpeechDat database has been collected and annotated by MATRA NORTEL COMMUNICATIONS. 40 speakers have been added to the original 5,000 speakers to fit the requirements of the database. This database is partitioned into 18 CDs, each of which comprises 300 speakers sessions (except for CD 4, with 100 speakers sessions). The speech databases made within the SpeechDat(II) project were validated by SPEX, the Netherlands, to assess their compliance with the SpeechDat format and content specifications.

The speech files are stored as sequences 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.

The following age distribution has been obtained: 215 speakers are below 16 years old, 2531 speakers are between 16 and 30, 1208 speakers are between 31 and 45, 910 speakers are between 46 and 60, and 176 speakers are over 60.

A pronunciation lexicon with a phonemic transcription in SAMPA is also included.

| **Price for ELRA members:** | | **Price for non members:** | |
|---|---|---|---|
| for research use: | 35,000 EURO | for research use: | 60,000 EURO |
| for commercial use: | 60,000 EURO | for commercial use: | 70,000 EURO |

## ELRA-S0077 Telephone Speech Data Collection for Czech

This database contains speech collected in Czech Republic during summer 1999. The collection was performed at the Institute of Radioelectronics of Brno University of Technology, Faculty of Electrical Engineering and Computer Sciences (VUT Brno) and at the Department of Circuit Theory of Czech Technical University in Prague, Faculty of Electrical Engineering (CVUT Prague) upon demand of Siemens AG, Corporate Technology, Munich. This database comprises telephone recordings from 1227 speakers (590 males and 637 females) recorded directly over the fixed telephone network using an ISDN interface.

Speech files are stored as sequences of 8bit 8 kHz A-law uncompressed speech samples. Each prompted utterance is stored within a separate file. Each speech file has an accompanying ASCII SAM label file according to the specifications of the SpeechDat project (URL: http://www.speechdat.com).

Corpus contents:
· connected digits (prompt sheet number, telephone number, credit card number),
· sequences of isolated digits (5 digits),
· answers to yes/no questions,
· common application words and phrases.

The following age distribution has been obtained:
- 36 speakers are below 16 years old,
- 537 speakers are between 16 and 30,
- 306 speakers are between 31 and 45,
- 259 speakers are between 46 and 60,
- 88 speakers are over 60,
- and 1 speaker whose age is unknown.

The transcription included in this database is an orthographic, lexical transcription with a few details that represent audible acoustic events (speech and non speech) present in the corresponding waveform files. SpeechDat conventions were used in this database.

| | |
|---|---|
| **Price for ELRA members:** | 10,000 EURO |
| **Price for non members:** | 15,000 EURO |

## ELRA-S0079 Finnish Speechdat(II) FDB-4000

The Finnish SpeechDat(II) FDB-4000 comprises 4000 Finnish speakers (1830 males, 2170 females) recorded over the Finnish fixed telephone network. The SpeechDat database has been collected and annotated by the Tampere University of Technology's Digital Media Institute. The FDB-4000 database is partitioned into 14 CDs, 13 CDs comprise 300 speakers sessions, the 14th comprises 100 speakers. The speech databases made within the SpeechDat(II) project were validated by SPEX, the Netherlands, to assess their compliance with the SpeechDat format and content specifications.

Speech samples are stored as sequences of 8-bit 8 kHz A-law. Each prompted utterance is stored in a separate file. Each signal file is accompanied by an ASCII SAM label file which contains the relevant descriptive information.

Each speaker uttered the following items:
· 1 isolated digit
· 1 sequence of 10 isolated digits
· 4 numbers: 1 sheet number (5 digits), 1 telephone number (9-10 digits), 1 credit card number (16 digits), 1 PIN code (6 digits)
· 1 currency money amount
· 1 natural number
· 3 dates: 1 spontaneous date (birthdate), 1 prompted date, 1 relative or general date expression
· 2 time phrases: 1 time of day (spontaneous), 1 time phrase
· 3 spelled words: 1 spontaneous own forename, 1 city name, 1 phonetically rich word
· 5 directory assistance names: 1 spontaneous own forename, 1 spontaneous city of growing up, 1 frequent city name, 1 frequent company name, 1 common forename surname
· 2 yes/no questions: 1 predominantly "yes" question, 1 predominantly "no" question
· 3 application words
· 1 word spotting phrase using an embedded application word
· 4 phonetically rich words
· 9 phonetically rich sentences

The following age distribution has been obtained: 545 speakers are below 16 years old, 1773 speakers are between 16 and 30, 980 speakers are between 31 and 45, 606 speakers are between 46 and 60, and 96 speakers are over 60.

A pronunciation lexicon with a phonemic transcription in SAMPA is also included.

| **Price for ELRA members:** | | **Price for non members:** | |
|---|---|---|---|
| for research use: | 30,000 EURO | for research use: | 45,000 EURO |
| for commercial use: | 40,000 EURO | for commercial use: | 50,000 EURO |

## ELRA-S0078 Finnish Speechdat(II) FDB-1000

This resource is a sub-set of ELRA-S0079 Finnish Speechdat(II) FDB-4000.

The Finnish SpeechDat(II) FDB-1000 comprises 1000 Finnish speakers (617 males, 383 females) recorded over the Finnish fixed telephone network. The SpeechDat database has been collected and annotated by the Tampere University of Technology's Digital Media Institute. The FDB-1000 database is partitioned into 4 CDs, 3 of which comprise 300 speakers sessions, and the fourth one 100 sessions.

Each speaker uttered the same items as for ELRA-S0079 Finnish Speechdat(II) FDB-4000.

The following age distribution has been obtained: 57 speakers are below 16 years old, 609 speakers are between 16 and 30, 223 speakers are between 31 and 45, 104 speakers are between 46 and 60, and 7 speakers are over 60.

| **Price for ELRA members:** | | **Price for non members:** | |
|---|---|---|---|
| for research use: | 9,000 EURO | for research use: | 22,000 EURO |
| for commercial use: | 18,000 EURO | for commercial use: | 25,000 EURO |

## ELRA-S0080 Finnish-Swedish Speechdat(II) FDB-1000

The Finnish-Swedish SpeechDat(II) FDB-1000 comprises 1000 Finnish speakers (455 males, 545 females) uttering speechdat items in the variant of Swedish spoken in Finland, recorded over the Finnish fixed telephone network. The SpeechDat database has been collected and annotated by the Tampere University of Technology's Digital Media Institute. The FDB-1000 database is partitioned into 4 CDs, 3 CDs comprise 300 speakers sessions, the 4th comprises 100 speakers. The speech databases made within the SpeechDat(II) project were validated by SPEX, the Netherlands, to assess their compliance with the SpeechDat format and content specifications.

Speech samples are stored as sequences of 8-bit 8 kHz A-law. Each prompted utterance is stored in a separate file. Each signal file is accompanied by an ASCII SAM label file which contains the relevant descriptive information.

Each speaker uttered the following items:
· 1 isolated digit
· 1 sequence of 10 isolated digits
· 4 numbers: 1 sheet number (5 digits), 1 telephone number (9-10 digits), 1 credit card number (16 digits), 1 PIN code (6 digits)
· 1 currency money amount
· 1 natural number
· 3 dates: 1 spontaneous date (birthdate), 1 prompted date, 1 relative or general date expression
· 2 time phrases: 1 time of day (spontaneous), 1 time phrase
· 3 spelled words: 1 spontaneous own forename, 1 city name, 1 phonetically rich word
· 5 directory assistance names: 1 spontaneous own forename, 1 spontaneous city of growing up, 1 frequent city name, 1 frequent company name, 1 common forename surname
· 2 yes/no questions: 1 predominantly "yes" question, 1 predominantly "no" question
· 6 application words
· 1 word spotting phrase using an embedded application word
· 4 phonetically rich words
· 9 phonetically rich sentences

The following age distribution has been obtained: 178 speakers are below 16 years old, 412 speakers are between 16 and 30, 216 speakers are between 31 and 45, 160 speakers are between 46 and 60, and 34 speakers are over 60.

A pronunciation lexicon with a phonemic transcription in SAMPA is also included.

| **Price for ELRA members:** | | **Price for non members:** | |
|---|---|---|---|
| for research use: | 9,000 EURO | for research use: | 22,000 EURO |
| for commercial use: | 18,000 EURO | for commercial use: | 25,000 EURO |

## ELRA-W0020 PAROLE French Corpus

The ELRA resource W0020 PAROLE French Corpus, described in the Newsletter Vol.4 n.4, is now available for distribution.

**Price for ELRA members (for research use only):** 1,540 EURO

**Price for non members (for research use only):** 4,300 EURO

## ELRA-W0022 ILSP/ELEFTHEROTYPIA Corpus (PAROLE Greek Corpus)

The ILSP/ELEFTHEROTYPIA Corpus contains approximately 3 million words classified and annotated according to the common core PAROLE encoding standard. Thus, each file is classified according to the parameters of Medium, Topic and Genre, and structurally annotated at paragraph level (CES Level 1). The format of the corpus is SGML files. The source of the files is the daily newspaper ELEFTHEROTYPIA.

A subset of the corpus (250,000 words) is morpho-syntactically tagged; all the words are also lemmatised and checked. For the morphosyntactic annotation of the corpus, a stepwise procedure consisting of the following four steps was used: automatic morphosyntactic annotation, automatic disambiguation, manual disambiguation and checking, conversion into the PAROLE format requirements. In certain texts, some passages are written in "katharevoussa", an older version of Greek; these passages are marked as "distinct" and have not been morpho-syntactically annotated.

The tagset used for the morphological annotation of the corpus is presented in the "*Addendum to TA - Encoding features and values for the morphological layer in the lexicon Merged Tags*" (P-WP1.1.-MEMO-ERLI-5).

| | |
|---|---|
| **Price for ELRA members (research use only):** | 850 EURO |
| **Price for non members (research use only):** | 1,275 EURO |

## ELRA-L0032 PAROLE Greek Lexicon

The PAROLE Greek lexicon has two layers, morphological and syntactic. It includes the most frequent words found in a 9 million word corpus, coded according to the PAROLE specifications.

The Morphological layer contains a total of 20149 Morphological units, of which 12042 are nouns (common and proper), 3014 verbs, 3405 adjectives, 106 numerals, 45 pronouns, 2 articles, 1396 adverbs, 48 adpositions, 51 conjunctions, 21 interjections, 19 "unique" categories.

The Syntactic layer contains 25092 Syntactic units, of which 14548 are nouns, 5397 verbs, 3558 adjectives, 1410 adverbs, 73 adpositions and 106 numerals.

This lexicon was constructed based on the following resources:

a. the ILSP Morphological Lexicon

b. the ILSP Corpus

| | |
|---|---|
| **Price for ELRA members (research use only):** | 3,400 EURO |
| **Price for non members (research use only):** | 5,100 EURO |

## Up-date of LantMark Lexica's prices

All prices are indicated in EURO.

*R: research use  C: commercial use*

### ELRA-L0004 Dutch lexicon, 64000 entries

| Prices for members | Prices for non members |
|---|---|
| R. 7,680 | R. 12,800 |
| C. 61,440 | C. 102,400 |

### ELRA-L0005 French lexicon, 50000 entries

| Prices for members | Prices for non members |
|---|---|
| R 6,000 | R. 10,000 |
| C. 48,000 | C. 80,000 |

### ELRA-M0004 Dutch-French lexicon

General and Specialised vocabularies for transfer

Entries: i) General Vocabulary (26 000), ii) Administrative (32 000), iii) Data processing (10 000).

| Prices for members | | Prices for non members | |
|---|---|---|---|
| R. | C. | R. | C. |
| i) 3,120 | i) 24,960 | i) 5,200 | i) 41,600 |
| ii) 3,840 | ii) 30,720 | ii) 6,400 | ii) 51,200 |
| iii) 1,200 | iii) 9,600 | iii) 2,000 | iii) 16,000 |

### ELRA-M0005 English-French lexicon, 33287 entries

| Prices for members | Prices for non members |
|---|---|
| R. 3,994.44 | R. 6,657.40 |
| C. 31,955.52 | C. 53,259.20 |

### ELRA-M0006 French-Dutch lexicon

General and Specialised vocabularies for transfer

Entries: i) General Vocabulary ii) (34 000), Administrative (18 000), iii) Data processing (10 000).

| Prices for members | | Prices for non members | |
|---|---|---|---|
| R. | C. | R. | C. |
| i) 4,080 | i) 32,640 | i) 6,800 | i) 54,400 |
| ii) 2,160 | ii) 17,280 | ii) 3,600 | ii) 28,800 |
| iii) 1,200 | iii) 9,600 | iii) 2,000 | iii) 16,000 |

### ELRA-M0007 French-English lexicon, 39453 entries

| Prices for members | Prices for non members |
|---|---|
| R. 4,734.40 | R. 7,890.60 |
| C. 37,874.90 | C. 63,124.80 |