The ELRA Newsletter

July - September 99

Vol.4 n.3

Contents

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

Editor in Chief:
Khalid Choukri

Editor:
Jeff Allen

Layout:
Audrey Mance
Valérie Mapelli

Contributors:
Daniel Brockmann
Xavier Garcia
Jeremy Peckham

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 or
WWW:
http://www.icp.grenet.fr/ELRA/home.html

Dear Members,

As we announced in previous issues of the ELRA Newsletter, the ELRA 1999 call for Language Resource (LR) Packaging and Production has led to 8 signed contracts for projects that are currently underway. All of these projects are expected to be completed during Spring 2000. In addition, ELDA has negotiated 3 other agreements to package modern French corpora, with the financial support of the French government. More details are at our Web site.

As for the overall LE4-8335 LRsPP project that concerns LR market monitoring and market segmentation, ELDA has sent the due set of deliverable reports to the European Commission. These reports include general statistics obtained from efforts conducted during the spring and summer of 1999. The ELRA Members' User Needs survey has resulted in a 40% reply rate thus far. We also sent out a summer 1999 User Needs survey to nearly 700 non-ELRA member LR users. Although all specific details (institution names and addresses, specific needs, etc) obtained from the questionnaires will remain confidential, ELRA expects to share some general statistics obtained from these surveys in upcoming newsletters, We encourage you to participate in these surveys if you have not yet already done so. Survey forms can be obtained from Jeff Allen (jeff@elda.fr).

During this quarter, ELDA submitted 2 project proposals for the September MLIS call and awaits responses from the EU.

With employees having a strong background in LR design, collection and implementation, ELDA is pleased to announce that it now offers a new "Language Resource collection service". Acting as a service provider, ELDA is prepared to work with institutions on a case-by-case basis in order to specify, collect and validate LRs that respond to specific needs. We would like to remind our members and partners who are/will be involved in EC-funded projects under the IST program that ELDA can provide the following services: legal and contractual assistance for negotiating a resource with a producer; information on other contractual or legal matters; advice on database design, collection, and validation procedures; LR identification service on other databases available or databases being developed.

Over the last few months ELDA has duplicated a large number of SpeechDat databases using its CD duplication platform. We encourage our partners to consider ELDA for future CD duplication needs.

In this issue of the ELRA newsletter, Jeremy Peckham's (Strategis Consulting) article provides a summary of various consumer areas that are and will continue to be affected by speech recognition technologies. Following Sharon O'Brien's article in Vol4 N2 on Translation Memory (TM) systems and LRs, this issue gives some additional perspectives with one paper on TRADOS products by Daniel Brockmann and one paper by Xavier Garcia on Atril Déjà Vu.

We wish to congratulate Joseph Mariani, one of ELRA's longstanding Vice Presidents, for the "Special Service Award" that he recently received from ESCA. This is described further in this issue. ELRA, wishing to stimulate high-quality contributions of papers at major conferences dealing with Human Language Technologies, offered a prize at Eurospeech99 for the best student paper addressing issues related to LRs. In addition, ELRA/ELDA presented a paper on its recent and on-going activities at Eurospeech99 and at the Cocosda Workshop.

ELRA is proud to welcome several new members since the publication of our last issue of the Newsletter: British Telecom Labs, UK; Hong Kong University of Science and Technology, Hong Kong; Universidad Europea de Madrid, Spain; Istituto Trentino di Cultura, Italy. In our "New Resources" section of this issue, we are glad to announce an updated list of EuroWordNet LRs. Full details on the PAROLE French Corpus are also provided. Danish and Swedish SpeechDat(II) LRs are also now available with further details available at the ELRA catalogue. The PAROLE Italian corpus and lexicon are expected to be made available soon. Please also note the correction for the Dutch PAROLE corpus and lexicon prices.

We would like to take the opportunity to welcome Audrey Mance who has recently joined ELDA as another one of our Technical Assistants.

Lastly, readers of the ELRA Newsletter in many countries will be deeply saddened to learn of the death of Ole Norling-Christensen this summer. We extend our condolences to family members and close colleagues who have been most affected by this loss, in particular to those who worked with and appreciated Ole within the PAROLE project.

Antonio Zampolli, President Khalid Choukri, CEO

Speech Technology - gaining ground

Jeremy Peckham, Strategis Consulting

Speech technology has come a long way in the last few years when judged by a number of different criteria such as capability, applications, market size and commercial interest. From a technology perspective a number of significant hurdles have been overcome, resulting in capabilities which, whilst still not being perfect, are sufficient to meet certain market needs. In particular, improved accuracy of speech recognition through training on vast quantities of data has opened up the telephony, desktop and consumer appliance markets. At all levels in the recognition and understanding process, whether at the acoustic-phonetic level or at the word level, data has been instrumental in improving performance.

More language capabilities have also resulted from increasing commercial interest in the global markets for speech technology. Interestingly these developments have also relied heavily on the availability of sufficient quantities of data. Given the huge cost involved in creating quality speech and text databases, the collaborative ventures of the last few years, both private and EC supported have been highly influential in broadening the language coverage of speech technology. In speech synthesis, despite the growing trend towards concatenative segment synthesis (including diphones and phones), progress has in some languages been slowed by the lack of detailed background work on letter to sound conversion rules. This has left the quality of some text to speech synthesis systems largely inadequate for public use.

Although much more research can inevitably be done on existing techniques one cannot help wondering whether current technology is reaching a plateau of capability, at least as far as the algorithms are concerned. Whilst this may be hotly debated, there is little doubt that much can be done in the area of usability engineering, taking today's limited capabilities and finding ways to deploy them that meet user expectations.

For years protagonists have played up the naturalness of speech as a means of human-machine communication. Even Bill Gates has now bought into this view and is spending significant dollars on research and investments in companies such as Lernout and Hauspie. To quote this captain of the software industry "I'd be so bold as to say that 10 years from now every personal computer will have been seeing, listening and learning" (CA-World 1998).

The compelling attraction of speech recognition is obvious - to enable people to talk to computers, dictate memos and access information over the telephone without touching the keyboard. However, when the spoken interface fails to work in the way that most users expect they become disenchanted, after all why should people be expected to learn to speak differently when the main point about the speech interface is its ease of use! This failure of technology to live up to the promise has been the single most critical factor in the slow progress in developing mass markets for speech technology. Success comes when the technology is embedded in applications which are sufficiently compelling and which work at a level perceived to be acceptable to the users.

Some of the users of the early voice dictation packages were highly motivated and put up with significant restrictions imposed by the limitations in technology. These early adopters were crucial to the development of the technology, allowing improvements to be made and important lessons to be learnt about what users want in the application. As dictation capabilities have improved and the price has plummeted, more and more business users are adopting the technology and are finding it productive, particularly where keyboarding skills are minimal. For experienced keyboard users, improvements in accuracy or retraining in working methods will probably still be required to avoid the technology getting in the way.

Dictation software has begun to make it into the mainstream PC market with many retailers now featuring one or more vendors products on prime shelf space. Whether this has been a good move or not long term remains to be seen, for what is becoming clear is that many casual purchasers of dictation software fare less well with it than those who have had some hand holding and training in installing and using the software. Details such as microphone positioning and quality of the sound card installation can all make a difference to performance. At prices as low as $40 though, perhaps few will complain when the product fails to live up to expectations.

In the telecommunications market however, the damage caused by a miss match in expectations is potentially huge in terms of customer loyalty and brand image. It is for this reason as well as other priorities that telephone companies have remained far more cautious in their adoption of speech technology. Despite this, many niche applications are being developed around the world by Industry leaders such as Philips and Nuance, providing stock quote services for small, closed user groups or airline information systems like Lufthansa's, where customers can self select whether they use the service. There are now an increasing number of services going on line covering a wide range of applications from parcel tracking, restaurant guides, frequent flyer information, home banking, train timetables and call routing.

Some of the most successful applications to date have been those where the consumer is largely unaware that automation was taking place. AT&T's collect call service in the USA is a case in point. Considerable development went into the creation of this application which, from the perspective of the caller, appears to only recognise two words - "yes" and "no". In reality a large number of variations in expression had to be handled.

Successful speech applications in the telecommunications world illustrate the need to carefully match what the technology can achieve with consumer expectations. It is little use applying spoken natural language dialogue technology in a mission critical scenario for a wide consumer user group without operator fall back. By the same token, requiring users to navigate a menu by voice commands violates the naturalness that we are so keen to promote. User expectations can of course be modified to be more favourable to today's capability under certain circumstances. Automated information services may be acceptable if the cost of using such services is less that the normal operator based service or the service is perceived as more convenient than other options such as touch-tone.

In some applications such as network based hands free dialling, speech recognition has mostly failed to consistently deliver the performance needed. The application has also often introduced call set up delays due to

the way the service has been implemented (nothing to do with speech recognition!) and it has required tedious training to set up the directory. Given these limitations the application has probably not delivered enough to the user. The personal assistant on the other hand started with too many features and instead of allowing users to navigate the service naturally, it required them to learn the commands needed to navigate menus - few seemed willing to do this. It is early days for a new generation of services created by companies like General Magic, but a better user interface coupled with features tailored to particular user groups will be key to acceptance and take up.

There has recently been a lot of interest by technology vendors in creating voice access to the web. This has come in different guises from the speech driven browser on a PC to telephony based applications which use standard HTML or Motorola's new mark-up language for voice, VoXTML

These initiatives, particularly the telephony ones pose a number of questions about the appropriateness of voice activation of web sites and the markets readiness to adopt such an approach. Are these initiatives simply an attempt by speech companies to mount the Internet bandwagon or is there something of more substance behind them?

The argument of the speech companies promoting access to the web by voice rather than PC is that it opens up the channel to people who don't have access to a PC. Where does this place the call centre that is currently the primary access channel for those who don't want to communicate electronically? Self-service IVR applications have grown up around the call centre to alleviate human agents from routine tasks and avoid lengthy call queues. Why then shift from a call centre centric approach to a web centric approach? Although much is made of the need to integrate web sites with the call centre to achieve better customer contact and relationship management, few companies have yet achieved this.

There do seem to me to be some potential problems and dangers with a web centric approach to voice interaction, illustrated by some of the demonstrations available to date from its promoters. Firstly natural language speech based IVR applications, if they are well designed exploit the advantages of speech and avoid users remembering long lists or navigating complicated menu structures. Web sites however are designed to be visually rich and require form filling, menu selection and browsing using visual queues that are well suited to the PC medium. The visual and tactile style of interaction does not always map well onto speech input and output and great care is needed to ensure that the web design does not drive the vocal/aural interface design. Some vendor's demos for example display a fundamental error of speech interface design by overloading the user with too many options at each branch of the menu. Where the input is the name of a film or a location, something which the user can be expected to know, long list recognition can work well. However, many web sites provide only limited lists of options but these are easy to scroll through or select from a screen display. To exploit the full advantage of speech as an interface will require a significant investment in dialogue engineering and a mapping of this interface onto the databases used by a web site.

Apart from the shared information and database which will eventually be the norm for call centres, there seem to be few other advantages to interacting with a web site by voice over traditional IVR approaches. There are perhaps some exceptions, where a web site has a relatively flat structure and the options are intuitive, adding "IVR" becomes very simple. At the end of the day I believe that web sites will merely be an alternative source of content for interactive spoken language dialogue systems and that significant work will still be required in developing the user interface to truly exploit the benefits of speech. In the meantime, in the crazy climate of over hyped Internet stocks, linking speech products to the Internet may help stock market valuations of speech companies and attract interest from investors, but will it create a real mass market?

Whilst progress has been made in technology capabilities and in finding good applications, there has also been much activity on the commercial front over the last few years with IPOs (Initial Public Offerings), large scale investments and some industry consolidation. Dragon announced its IPO and in the process revealed turnover for 1998 in excess of $70m, although it has since withdrawn the IPO. Philips has reached agreement to acquire VCS, which itself had absorbed Scott Instruments, VPC and Pure Speech. Lernout and Hauspie have also acquired a number of speech companies such as Kurtzweil and Centigram as well as other language-based organisations including translation services. Companies such as Intel and Microsoft, as well as Venture Capitalists in the USA and Europe have invested well over $500m in the speech industry in the last couple of years.

What then does the future hold for the industry and for research? Speech technology is clearly here to stay and will begin to pervade more and more of our lives and activities. The increase in efficiency of algorithms coupled with reducing costs of memory and rising processor power will mean that speech technology can be realistically incorporated into more consumer products from telephones to car dashboards and palm top computers. Pressures on the call centre and increasing volumes of callers will drive further adoption of Interactive Voice Response regardless of whether it will be coupled to the web site. Self service applications will dominate until companies have more confidence in the technology and even then, careful integration with agent based interaction will be required to deliver a high quality service. Network based services have yet to really establish themselves in the mainstream, largely a combination of technology limitations and poor application conception. Variants of the Personal Assistant concept are still very much at the trial stage and the jury is still out on how widely this concept will be adopted. The slowness of large PTOs to seriously take up speech technology make well result in their being overtaken by device orientated applications such as voice activated 'smart' phones and palm tops.

Ultimately, speech technology will find its way into the operating system of PC's, mobile computers and Network based systems. The availability of the core technology at low cost in the operating system though, will not itself ensure a mass market. Much work still remains to be done on usability and application integration. Tools for rapid development of spoken language applications are still in their infancy and these together with the development of new ideas in algorithms and architectures should keep the research community busy for some time to come!

Is there a breakthrough on the horizon? In over twenty years of involvement in the industry I can say that we have come a long way both technically and commercially. Much of this progress has been slow, required painstaking attention to detail and commitment to the end game. I believe that the future will be no different than the past. Progress will be incremental but at some point, talking and listening to machines will be an every day occurrence for many people and speech technology will be as important as the keyboard is today.

Jeremy Peckham
Strategis Consulting, 49 Hinton Road
Fulbourn, Cambridge CB1 5DZ, UK
Tel.: +44 1223 500844
Fax: +44 1223 501974
Email: JBPeckham@aol.com

Beyond "fuzzy matching": The Déjà Vu approach to reusing Language Resources

Xavier Garcia, Amperstand Traducció Automática

Background

During the last 10 years a new approach to efficient, high-quality human translation has been attempted: reuse of translation language resources or aligned translation corpora. Some TM companies have developed software programs capable of storing source texts and target text in databases allowing its reuse, based on simple algorithms that perform so-called "fuzzy matching".

This article explains the limitations and shortcomings of this matching approach and describes a solution that is currently in use (and still developing further) in Atril Software's Déjà Vu translation platform, a translation software suite based on more in-depth reuse of LRs and on the principle of controlled subsentence translation storing and reuse.

LR reuse: a classic problem

Traditional Translation Memory systems are not users but are generators of language resources. Surely, they do "reuse" language resources, but only those created by the same users; TMs are seldom of any use to other users.

Why? Because of the way of searching for similar translations in current TM systems, based on simple string-comparison algorithms and on the definition of a full sentence as a translation unit.

This principle of a simple comparison of full-sentence translation units is not consistent with specific-purpose language nature. Indeed, in those situations language tends to repeat itself -but in smaller units.

Anyone translating legal documents (or having to wrestle with EC Calls for Proposals, for that matter) knows that legal language (Legalese, a domain in itself) is a boring, repetitive language, yet "fuzzy match" analysis will show little full-sentence repetition benchmarks, thus making legal language corpora databases unusable to anyone except the translator who created them.

On top of that, the fuzzy match approach is not a guarantee of translation quality (i.e., consistency, in today's ISO-9000 world): ensuring consistent reuse of identical or almost-identical sentences does not guarantee consistent reuse of the embedded subsentences (or "chunks"), often amounting to 60% of the total corpus weight.

The Déjà Vu solution

Atril Software is implementing in its Déjà Vu TM tool more flexible storing and retrieving algorithms that allow 3 kinds of new processes to process and reuse, not only translation units, but language at large:

First, the user can independently store arbitrary multi-word units (MWU). These units (a concept that goes beyond the traditional terminology entry), are then used by Déjà Vu for pretranslating long sentences, even (and specially) when the whole sentence is not found in the corpus database, not even in a fuzzily similar way.

Second, Déjà Vu itself can use the combination of full-sentence and subsentence translation units for adding-and-substracting different-length units. It extracts new subsentence translation units not previously found in the translation database as such. And it finally combines them in order to find a very usable, but not perfect, translation proposal.

Third (a power feature still under development), Déjà Vu will soon be able to independently search the whole content of the translation database in order to assemble translations of almost entirely new sentences, making it up of bits and pieces spread out all over the memory. In other words, using language to produce language.This is very similar to Example-Based Machine Translation.

A low-level example of this third new technology can be found in the "Learn" function of Déjà Vu: the current version (2.3.55) of the program can already guess the translation of a word or group of words based only on statistical "data mining" of the corpus database. The bigger (and better) the memory, the higher the chances of successful matches. The automation of this process for the whole translation sequence is the last milestone for our developers to reach.

The future: the final emergence of a Language Resource market

Developments like the one we have described above encourages the on-going need for language resources: they will become much more usable --not just "very interesting" to study. In the near future, any English-into-French legal translator, for instance, will for the first time be able to buy, sell or exchange translation corpus databases with his or her colleagues as a basic but useful-in-practice language resource, because it will be usable by third parties. That is an important improvement for the field!

For this same reason, the economic implications of the practical reusability of corpora suggest a financial viability of a corpora-creating industry, a field needing much funding, as a new activity powered by the emerging (and already powerful) language industry.

Companies like Atril and many others will welcome and encourage the development of such a market, one which opens new perspectives and development paths for the cooperation between research and industrial players.

About Déjà Vu

Please see Atril's website at www.atril.com for more information on Dejà Vu.

Xavier Garcia
Amperstand Traducció Automática
Travessera de Gracia, 73, 1-7
08006 Barcelona, Spain
Tel. : + 34 93 415 9990
Fax : + 34 93 416 1862
Email : xavi@amperstandsl.com

Conference announcer
SECOND INTERNATIONAL CONFERENCE O
LRE

ATHENS, GREECE

Ho
The Institute for Language and Spe
The National Technical U

The Second International Conference on Language Resources and Evaluation has been initiated by ELRA and is organised in cooperation with other Associations and Consortia, including ACL, ALLC, COCOSDA, ORIENTAL COCOSDA, EAFT, EAGLES, EDR, ELSNET, ESCA, EURALEX, FRANCIL, LDC, PAROLE, TELRI, etc., and with major national and international organisations, including the European Commission - DG XIII, ARPA, NSF, the IC/863 HTRDP Project (China), the National Natural Science Foundation of China, the ICSP Permanent Committee (Korea), The Natural Language Technical committee of JEIDA (Japan), and the Japanese Project for International Coordination in Corpora, Assessment and Labelling. Cooperation and support from other institutions is currently being sought.

CONFERENCE AIMS

In the framework of the Information Society, the pervasive character of Human Language Technologies (HLT) and their relevance to practically all the fields of Information Society Technologies (IST) has been widely recognised. Two issues are currently considered particularly relevant: 1) the availability of language resources and 2) the methods for the evaluation of resources, technologies and products. Substantial mutual benefits can be expected from addressing these issues through international cooperation. The term language resources (LR) refers to sets of language data and descriptions in machine readable form, used specifically for building and evaluating natural language and speech algorithms or systems, for software localisation industries and language services, for language enabled information and communication services, for electronic commerce, electronic publishing, language studies, subject-area specialists and end users. Examples of language resources are written and spoken corpora, computational lexica, grammars, terminology databases, and basic software tools for the acquisition, preparation, collection, management, customisation and use of these and other resources.

The relevance of evaluation for Language Engineering is increasingly recognised. This involves assessment of the state of the art for a given technology, measuring the progress achieved within a programme, comparing different approaches to a given problem and choosing the best solution, knowing its advantages and drawbacks, assessment of the availability of technologies for a given application, product benchmarking, and assessment of user satisfaction. Language engineering and R&D in language technologies have made important advances in the recent past in various aspects of both written and spoken language processing. Although the evaluation paradigm has been studied and used in large national and international programmes, including the US ARPA HLT programme, the EU LE programme under R&D framework programmes, the Francophone Aupelf-Uref programme and others, and in the localisation industry (LISA and LRC), it is still subject to substantial unresolved basic research problems. The aim of this conference is to provide an overview of the state of the art, to discuss problems and opportunities, and to exchange information regarding ongoing and planned activities, language resources and their applications. We also intend to discuss evaluation methodologies and demonstrate evaluation tools, and explore possibilities and promote initiatives for international cooperation in the areas mentioned above.

CONFERENCE TOPICS

The following non-exhaustive list gives some examples of topics which could be addressed by papers submitted to the Conference:

1. Issues in the design, construction and use of Language Resources (LR) -theoretical & best practices

* Guidelines, standards, specifications, and models for LR
* Organisational issues in the construction, distribution, and use of LR
* Methods, tools, procedures for the acquisition, creation, annotation, management, access, distribution, and use of LR
* Legal aspects and problems in the construction, access, and use of LR
* Availability and use of generic vs. task/domain specific LR
* Methods for the extraction and acquisition of knowledge (e.g. terms, lexical information, language modelling) from LR
* Monolingual and multilingual LR
* Multimodal and multimedia LR
* LR and the needs/opportunities of the emerging multimedia cultural industry
* Industrial production and use of LR
* Integration of various modalities in LR (spoken, visual, gestual, textual)
* Exploitation of LR in different types of applications (language technology, information retrieval, vocal interfaces, electronic commerce, etc.)
* Industrial LR requirements and the community's response
* Analysis of user needs for LR
* Mechanisms of LR distribution and marketing
* Economics of LR
* Customisation and use of LR

ment and call for papers
N LANGUAGE RESOURCES AND EVALUATION
C-2000

, 31 MAY- 2 JUNE 2000

sted by
eech Processing (ILSP), Athens, Greece
University of Athens, Greece

* Research issues relevant for LR

2. Issues in Human Language Technologies evaluation

* Evaluation, validation, quality assurance of LR
* Benchmarking of systems and products; resources for benchmarking and evaluation
* Evaluation in written language processing (text retrieval, terminology extraction, message understanding, text alignment, machine translation, morphosyntactic tagging, parsing, semantic tagging, word sense disambiguation, text understanding, summarisation, localisation, etc.)
* Evaluation in spoken language processing (speech recognition and understanding, voice dictation, oral dialog, speech synthesis, speech coding, speaker and language recognition, etc.)
* Evaluation of document processing (document recognition, on-line and off-line machine and hand-written character recognition, etc.)
* Evaluation of (multimedia) document retrieval and search systems
* Evaluation of multimodal systems
* Qualitative and perceptive evaluation
* Evaluation of products and applications
* Blackbox, glassbox and diagnostic evaluation of systems
* Situated evaluation of applications
* Evaluation methodologies, protocols and measures
* From evaluation to standardisation of LR
* Research issues relevant to evaluation

3. General issues

* National and international activities and projects
* LR and the needs/opportunities of the emerging multimedia cultural industry
* Priorities, perspectives, strategies in the field of LR national and international policies
* Needs, possibilities, forms, initiatives of/for international cooperation

PROGRAM

The Scientific Programme will include invited talks, presentations of accepted papers, poster sessions, referenced demonstrations and panels.

FORMAT FOR ABSTRACT SUBMISSION

Submission of summaries for proposed papers and posters should consist of about 800 words. Demonstrations of LR and related tools will be reviewed as well. Please send an outline of about 400 words. If a demo is connected to a paper, please attach the outline to the paper summary.

A limited number of panels is foreseen. Proposals are welcome and will be reviewed. Please send a brief description, including an outline of the intended structure (topic, organiser, panel moderator (if different), tentative list of panellists). All submissions should include a separate title page, providing the following information: the type of proposal (paper or poster, demo, paper plus demo, panel); the title to be printed in the programme of the Conference; names and affiliations of the authors or proposers; the full address of the first author (or a contact person), including phone, fax, email, URL; the required facilities (overhead projector, data display; other hardware, platforms, Internet connections, etc.); and 5 keywords. All submissions will be reviewed by the Scientific Committee.

Electronic submission

Electronic submission of abstracts should be in ASCII file format.
This file should be sent to:
lrec@ilc.pi.cnr.it
Attn: Antonio Zampolli
LREC-2000 chairman

Submission in hard copy

You may also submit hard copies.
Please send five hard copies to:
Antonio Zampolli LREC-2000 chairman
Istituto di Linguistica Computazionale del CNR
via della Faggiola, 32 - 56126, Pisa, ITALY

CONFERENCE PROCEEDINGS

All the speakers accepted at the Conference (papers and posters) will be requested to provide the final version of their text for the CONFERENCE PROCEEDINGS by the 2nd of April 2000 (the instructions for the formatting of the final version of the text will be sent to the authors together with the notification of the acceptance, on the 2nd of February 2000). All the registered Conference participants will receive one copy of the proceedings on their arrival at the Conference Secretariat.

DEMONSTRATIONS & INTERNET FACILITIES

Internet connections and various computer platforms and facilities will be available at the Conference site. In addition to referenced demos concerning LR and related tools, it will be possible to run unreferenced demos of language engineering products, systems and tools. Those interested should contact the organiser of the demonstrations, Mr. S. Piperidis directly.

IMPORTANT DATES

Submission of proposals for papers, posters, referenced demos, panels and workshops: 20 Nov. 1999
Notification of acceptance of workshop and panel proposals: 10 Dec. 1999
Notification of acceptance of papers, posters, referenced demos: 2 Feb. 2000
Final version of the articles for the proceedings: 2 Apr. 2000

WORKSHOPS

Pre-Conference Workshops will be organised on the 29th and 30th of May and post-Conference Workshops on the 3rd and 4th of June. Proposals for workshops should be sent to Prof. A. Zampolli (see address below), be two to three pages in length and contain:

* A brief technical description of the specific technical issues that the workshop will address.
* The reasons why the workshop is of interest at the moment.
* The names, postal address, phone and fax numbers and email addresses of the Workshop Organising
* Committee, which should consist of at least three people knowledgeable in the field but not all from the same institution.
* The name of one member of the Workshop Organising Committee who is designated as the contact person.
* A schedule for organising the workshop and a preliminary agenda.
* A summary of the intended workshop Call for Participation.
* A list of audio-visual or technical requirements and any special room requirements.

The workshop proposers will be responsible for the organisational aspects (e.g. Workshop Call preparation and distribution, review of papers, notification of acceptance, etc.). Further details will be sent to the proposers.

WORKSHOP PROCEEDINGS

Each Workshop coordinator will collect the texts for the Workshop proceedings.
The subscription fees to a Workshop include a copy of the Workshop proceedings, which will be available at the Secretariat of the Conference.

CONSORTIA AND PROJECT MEETINGS

Consortia or projects wishing to take this opportunity for organising meetings, should contact the Conference Secretariat for assistance in arranging meeting facilities.

CONFERENCE ADDRESSES

Ms. Despina Scutari - Secretariat of the LREC-2000 Conference, general information
Institute for Language and Speech Processing
6, Artemidos & Epidavrou Str.
15125 Marousi Athens, Greece
Tel: +301 6800959
Fax: +301 6856794
E-mail: LREC2000@ilsp.gr

Mr. Stelios Piperidis - Demonstration organiser
Institute for Language and Speech Processing
6, Artemidos & Epidavrou Str.
15125 Marousi
Athens, Greece
Tel: +301 6800959
Fax: +301 6854270
E-mail: spip@ilsp.gr

Ms. Elsa Liakakou - Information on travel, accommodation and general information on Athens
MOEL
36, Eleon str, 14564, Nea Kifissia, Greece
Tel: +301 6203625
Fax: +301 8078342
E-mail: liagramo@internet.gr

CONFERENCE PROGRAMME COMMITTEE

Nicoletta Calzolari, Istituto di Linguistica Computazionale, Pisa, Italy
George Carayannis, Institute for Language and Speech Processing
Khalid Choukri, ELRA, Paris, France
Harald Höge, Siemens, Munich, Germany
Bente Maegaard, CST, Copenhagen, Denmark
Joseph Mariani, LIMSI-CNRS, Orsay, France
Antonio Zampolli, University of Pisa, Pisa, Italy (Conference chair)

INTERNATIONAL ADVISORY COMMITTEE

Sture Allen, professor, former permanent secretary of the Swedish Academy, Sweden
Souguil Ann, Seoul National University, Korea
Roberto Cencioni, Commission of the EU, DGXIII, Luxembourg
Zhiwei Feng, The State Language Commission of China, Beijing, China
Emm. G. Fragoulis, Secretary General for Research and Technology, Athens, Greece
Hiroya Fujisaki, Science University of Tokyo, Japan
Angel Martin Municio, President of the Real Academia de Ciencias, Madrid, Spain
Mark Maybury, MITRE Corporation, Boston, USA
Bernard Quemada, Conseil Supérieur de la Langue Française, Paris, France
Gary Strong, NSF & ARPA, Washington, D.C., USA
Piet G.J. Van Sterkenburg, International Permanent Committee of Linguists, Leiden, The Netherlands
Jialu Zhang, Academia Sinica, Institute of Acoustics, Beijing, China

LREC-2000 EXHIBITION

An exhibit area will be made available at LREC-2000. This is open to companies and projects wishing to promote, present and demonstrate their HLT products and prototypes to a wide range of experts and representatives from all over the world who will be participating at the conference. Please note that the exhibits of HLT products and prototypes are different from LR and system demonstrations accepted for presentation within the conference. The exhibits will run in parallel with the Conference for 3 days and the exhibit hall will be located near the general conference rooms.

LREC-98, in Granada, had over 197 papers and posters presented, with about 510 registered participants from over 38 different countries from all the continents. Among these, the largest group came from Spain (81 participants), followed by France (75), USA (73), Germany (47), UK (43) and Italy (41). Registered participants belonged to over 325 different organisations, out of which there were 115 industrial organisations and 210 academic institutions (universities, research centers). We therefore expect the exhibits at LREC-2000 to have a large audience.

For more information, please contact the ELDA office at: choukri@elda.fr

Translation Memories as True Databases: Present and Future

Daniel Brockmann, Trados GmbH

Today's translation memory (TM) systems can be regarded as highly specialised database and dedicated front-end applications which are optimised towards the processing of linguistic units - typically sentences - together with their translations. A highly efficient retrieval process in terms of speed, transparency, and quality, as well as a versatile and user-friendly translation front-end make up the key aspects of this optimisation. Moreover, since a TM system's core feature is translation re-use, the formatting aspects of the sentences are of crucial importance. Advanced features in this area include persistent sentence formatting (such as font information), formatting re-use across file formats (e.g. RTF vs. HTML), and automatic context-sensitive adaptation of formatting across documents. A TM system should also support typical advanced word processor features such as linguistic sub-units - footnotes, cross-references, and index entries - in a user-friendly manner. This article concentrates on some of these key features as they are implemented in the TRADOS Translator's Workbench.

What is a Translation Memory?

While the translator works, a TM system such as TRADOS Translator's Workbench dynamically builds a database that "remembers" all translations together with their source-language equivalents. Such a pair of source and target sentences is referred to as a *translation unit*. The database, itself referred to as *translation memory*, stores all translation units along with additional information such as administrative and user-defined data. In this process, special linguistic access structures are created to allow Translator's Workbench to find identical or similar sentences as rapidly as possible.

When the system encounters a sentence that has already occurred, it automatically retrieves the corresponding target-language sentence from the TM and presents this as the translation suggestion to the user. Unfortunately, 100% identical sentences don't turn up as often as one might hope. Much more common are sentences that have been slightly changed, for instance with a new product name or a different performance statistic. To find sentences that are only similar to each other, the computational linguists at TRADOS have over the last years refined what is referred to as *linguistic fuzzy matching.*

Linguistic Fuzzy Matching

Computers generally search for exact matches. Fuzzy matching is a technique for finding data that has only a certain degree of similarity to the search argument. In Translator's Workbench, this means that sentences are found in the TM even if they are only partially similar, and not necessarily identical, to others that have already been translated and exist in the database.

Small differences can alter the meaning of a sentence considerably. On the other hand, sometimes a sentence can be very similar to another in spite of more significant changes. To quickly find close matches with meaningful content, TRADOS Translator's Workbench uses an artificial neural network. A new sentence is matched against the ones already present in the neural network. Linguistic processing is carried out in the network to find the sentence in the translation memory that contains the fewest number of changes. This sentence in the TM is then selected as the "best match". However, other sentences that are less similar to the search sentence are not entirely discarded. Translator's Workbench adds all of them to the list of matches, thus allowing the translator to choose among several possibilities.

The translator will probably have to make minor modifications to the translation that is chosen as the "best match". To simplify this process, the system uses a colouring scheme to both a) signal the match quality and b) highlight the differences between the current sentence and the sentence from the translation memory. For instance, inserted words are displayed in grey, and changed words appear in purple.

During the fuzzy-matching process, Translator's Workbench calculates a percentage for each match, referred to as the match value, which expresses the degree of similarity between the search sentence and its counterpart in the TM. The higher the match value, the more similar the sentences are. A match value of 100% denotes what is referred to as a perfect match. To calculate the percent of similarity, the fuzzy-matching algorithm has to determine which portions of a sentence have changed, that is, which words and sentence parts were exchanged, deleted, inserted, or moved. Translators can then use this information in order to adapt the suggested translation as quickly as possible. They can set a minimum value, based on the fuzzy-match percentage calculation, that must be achieved for the system to suggest a translation. All sentences below this threshold are not processed and are thus translated manually.

Let's take a look at a few examples to clarify what has been said above. In the following instance, Translator's Workbench has found two matches for the new source sentence *What exactly is a translation memory?* In the "best match", valued at 86%, only the adverb has changed *(exactly vs. precisely)*. Notice the yellow colouring to highlight the changes *(Figure 1; editor's note: purple in our layout)*.

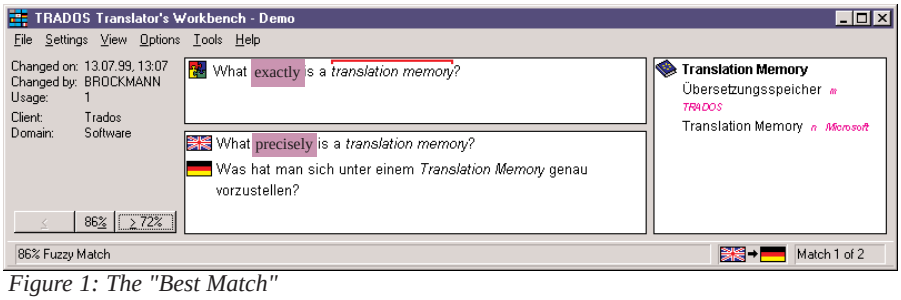


Figure 1: The "Best Match"

In the second match, valued at 72%, there are more significant changes. The sentence from the TM does not contain any adverb. It has been phrased slightly differently, which accounts for the lower match value*(Figure 2)*.

Still, the translation of the second match may actually be closer to what the user would like to put into their new translation. By presenting all possibilities to the user, Translator's Workbench leaves the choice to the user as to what match fits best into the current translation context.

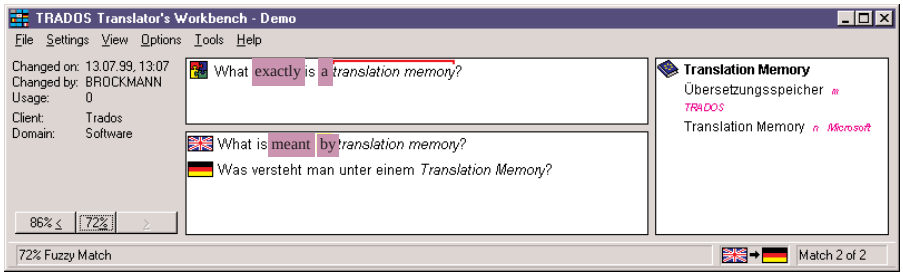


Figure 2: The Second Match

Fuzzy Terminology Recognition

Most modern computer-assisted translation systems not only feature a TM database with translation units, but also a terminology database with terms and additional information in several languages. In the case of Translator's Workbench, the terminology component is best known as TRADOS MultiTerm '95 Plus.

Each new sentence is not only matched against the translation memory, but also against the MultiTerm database to find any known terms. During this analysis, Translator's Workbench highlights all found terms in the source text and displays them in a separate window. A keystroke or mouse click then pastes the translation of the term into the document. The terminology matching, referred to as *active terminology recognition*, also works with a fuzzy-matching algorithm. As a consequence, it not only finds morphologically reduced forms, for example base forms of verbs, but also root forms of compound words, even if the elements of these compound words are spread over the sentence. Consider the following example:

One of the companies located on the Danube river produces steamboats.

In this example, Translator's Workbench will not only find the entries for company, locate, and produce, reducing the inflected forms in the sentence to the root form as stored in the MultiTerm database. It will also find the entry for Danube steamboat, although this compound does not occur as such in the sentence *(Figure 3)*.

Bilingual Concordance Searching

In addition to matching whole sentences and terms, Translator's Workbench also lets the user search for any text fragments in the translation memory. This feature is referred to as *bilingual concordance searching.* After selecting any text in the document and clicking on the mouse button, the translator can see all sentences in the translation memory containing the selected text fragment, along with their translation equivalents. This allows the user him or her to quickly see the searched text part in context, together with the appropriate translations. In the example below*(Figure 4)*, the translator has looked for the word *available* in the translation memory, which has three different German translations, depending on the context .

Translation and Automatic Substitution of Variable Elements

To further increase the overall translation throughput and improve the quality of fuzzy-matching, the current generation of Translator's Workbench (version 2.x) contains an advanced tokeniser that goes beyond the "standard" detection of word and sentence boundaries and also recognise and classify a wide range of so-called *variable elements*. Such elements can be numbers, acronyms, dates, time, measurements or members of user-defined word lists. One area of application where this advanced tokenisation is highly useful is the automatic adaptation to local data formats: Translator's Workbench "localises" the format of the elements as appropriate so that they appear in their correct form in the translation. For instance, in English-German translation, Translator's Workbench can adapt the English number format to the requirements of the German language, replacing the digit grouping symbol (",") and decimal symbol (".") with their German counterparts, which have to appear exactly the other way round.

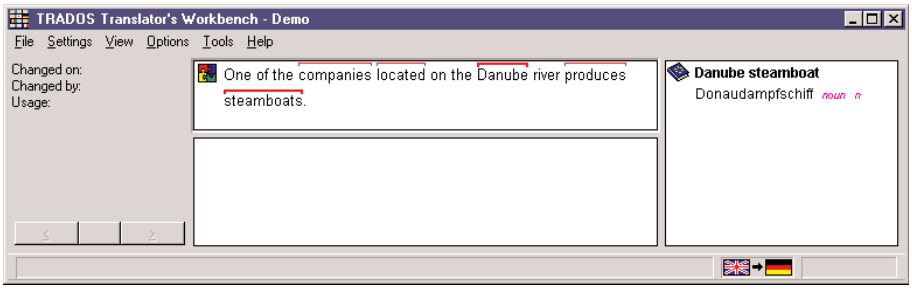


Figure 3: Active, Fuzzy Terminology Recognition

The use of information gained through advanced tokenisation is not limited to automatic localisation. In the following example *(Figure 5)*, Translator's Workbench recognises the acronym *DAX* and two numbers as variable elements. As a visual aid for the translator, the system places a bracketed line under them.

When transferring the numbers into the translation, Translator's Workbench will change their format so that they appear in the correct German form: They will be "localised" as 4.919,60 and 4.936,32.

In the translation memory, all variable elements appear in an abstract form. In the above example, after translation, the translation memory will contain the sentence The {ACRONYM} index hovered between {NUMBER} and {NUMBER}. This mechanism allows Translator's Workbench to automatically replace the variable elements, even if they have changed in a new sentence. For instance, Translator's Workbench will be able to automatically translate the sentence, say, The XETRA index hovered between 3,687.80 and 3,699.48 points, into German, although both the acronym and the numbers have changed with regard to the first example above.

Using TRADOS Tools in NLP Research and Development

Taking the above-described feature set into account, it is apparent that both MultiTerm and Translator's Workbench lend themselves readily to being used as standard applications for storing, manipulating, editing and using lexical and linguistic resources. MultiTerm, for example, provides open and well-documented interfaces for importing and exporting multilingual lexical and terminological data and can thus be used as a resource repository with additional fuzzy-matching capabilities and flexible database schemes. Translator's Workbench goes even further since its functionality is accessible via an OLE/COM-based API. This allows Translator's Workbench to be cross-linked smoothly with (NLP) applications that need to make use of bilingual aligned corpora through the powerful searching and maintenance features of Translator's Workbench.

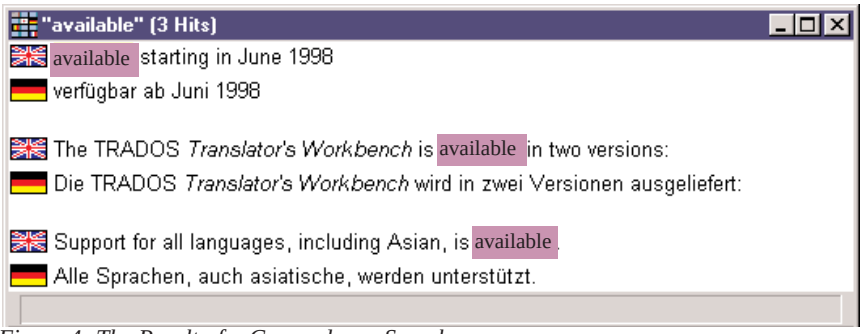


Figure 4: The Result of a Concordance Search

Future Trends in Translation Memory Technology

We see four major trends for the future development of translation memory technology. First of all, as European projects such as Otelo (http://www.otelo.lu) have shown, there is a growing demand for combining translation memory systems with machine translation engines to further increase translation productivity. The existing version of Translator's Workbench already features several possibilities to exchange translation information with machine translation systems. However, there is still room for improvement in some areas, e.g. for automating the combined translation memory/machine translation workflow, preserving all formatting, etc.

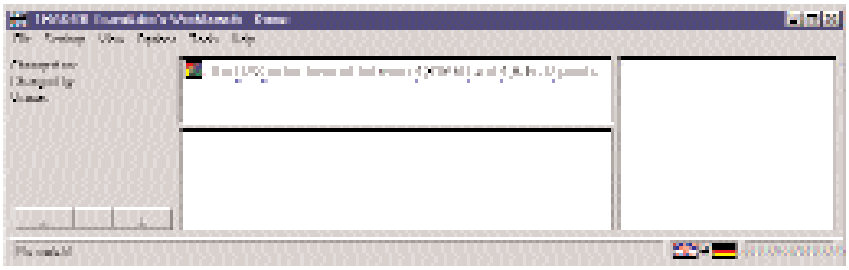


Figure 5: Variable Elements
(an Acronym and Two Numbers in this Example)

Secondly, the translation memory exchange format, abbreviated as TMX, will be fully implemented in all major translation memory systems in the near future in order to allow easier data exchange between the different products. More information on TMX is available at http://www.lisa.org/tmx.

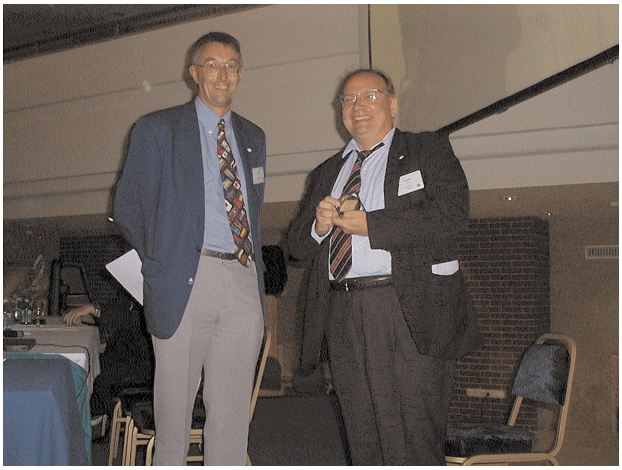
Thirdly, as more robust and efficient methods of NLP technology are being developed and seem to gain industry-level applicability and stability, new levels of linguistic processing can be integrated into TM systems. This will improve the "recycling rate" which can be achieved with translation memories and will increase the overall translation efficiency. Broad-coverage and robust methods are prerequisites for this integration, as well as their availability for a vast number of languages.

Finally, the translation process and the often massive concomitant costs it involves, will have an increasing impact on the design and authoring process of documentation. Translation memory applications and machine translation systems will make the step out of the language departments and will become one facet of the integration of the whole documentation workflow, including authoring, editing, and proof-reading, as well as translation and localisation.

Daniel Brockmann
TRADOS GmbH
Christophstr. 7
D-70178 Stuttgart
Tel. : + 49 (0711) 16877-50
Fax : +49 (0711) 16877-50
E-mail : daniel@trados.com

Special Service Medal

Joseph Mariani receives a special service award from ESCA



Last September, ESCA awarded a unique 'Special Service Medal' to ELRA's Vice-president Joseph Mariani, of LIMSI, for his key role as first President of the Association and for the establishment of the EUROSPEECH series of conferences.

The story actually starts back in May 1987, in Denmark, at the conference on speech technology organized at Jutland Telephone by the Danish Teletechnical society, and supported by the European Commission, which suggested the idea of starting a European Association on Speech Communication. The European Speech Communication Association was launched during a conference in Brussels on 26 February, 1988. René Carré was the chairman and Joseph Mariani elected as president of ESCA, which had the support of the Institute of Acoustics and of the French Acoustical Society. One of the first projects was to organize a conference, Eurospeech, which was to alternate with the International Conference on Spoken Language Processing (ICSLP) in Japan. The largest part of Joseph Mariani's mission ended at the Eurospeech'93, when he passed on the chairmanship to Louis Pols.

While being awarded this medal, Joseph Mariani mentioned people who provided major contributions to ESCA, in particular René Carré, Louis Pols, not to forget Max Wajskop and Christian Benoît, who are no more with us.

In the next issue, Joseph Mariani will present an article on the state of the art of speech technologies.

Joseph-Jean Mariani
LIMSI-CNRS
BP 133 91403 Orsay Cédex
France
Tel.: + 33 1 69 85 80 85
Fax: + 33 1 69 85 80 88
Email: mariani@limsi.fr

New Resources

EUROWORDNET

Following the announcement of the EuroWordNet databases in the last issue of the ELRA Newsletter (Vol.4 N.2), we are happy to announce that the list of EuroWordNet languages has grown. The following wordnets are now available via ELRA:

ELRA ref.	Language	Synsets	Word Meanings	Language Internal Relations	Equi-valence Relations
ELRA-M0015	English	16361	40588	42140	0
ELRA-M0016	Dutch	44015	70201	111639	53448
ELRA-M0017	Spanish	23370	50526	55163	21236
ELRA-M0018	Italian	40428	48499	117068	71789
ELRA-M0019	German	15132	20453	34818	16347
ELRA-M0020	French	22745	32809	49494	22730
ELRA-M0021	Czech	12824	19949	26259	12824
ELRA-M0022	Estonian	7678	13839	16318	9004

The prices are based on the number of synsets in each wordnet and differ for the kind of usage and ELRA-membership.

ELRA-W0020 PAROLE French Corpus

The PAROLE French corpus contains 20 093 099 words. The corpus consists of the following data:

- Miscellaneous: Data provided by ELRA (CRATER, MLCC Multilingual and Parallel Corpora) 2 025 964 words
- Periodicals: CNRS Info, Hermès 942 963 words
- Books: CNRS Editions 3 267 409 words
- Newspapers: Le Monde, provided by ELRA 13 856 763 words

1. Newspapers: 14 million words were extracted from complete issues of years 1987, 1989, 1991, 1993 and 1995 of Le Monde newspaper. 241 484 words, from 7 issues of Le Monde of September 1987, have been extracted, and POS-tagged automatically. Each article consists of a complete item - header - according to the directives of the TEI (Text Encoding Initiative). Le Monde original markups were changed into classication features, so that extracting articles of different topics is possible.

2. Periodicals:

- HERMES: Issues 15 to 22 have been used (134 articles, one Word file per article). The data have been converted from Word to RTF (Rich Text Format) and then, via a translator, from RTF to HTML. The conversion from HTML to the PAROLE format was made thanks to flex programs. The result for each article is: one "header" file which contains information on the author and the article id, and one "body" file which contains the article itself. A perl script is creating the final file from both "header" and "body".
- CNRS-Infos: The data come from the CNRS-Infos Web site (http://www.cnrs.fr/Cnrspresse/cnrsinfo.html). Each file has been as follows: cleaning the HTML header, extracting a summary, cleaning of HTML markups, translation to the PAROLE format, creation of the "header" and the "body" files (see Hermes). Like Hermes files, a perl script is creating the final file from both "header" and "body".

3. Books: All books were provided on CD-ROM as Xpress files, each book having its own structure. Therefore, each book have been considered separately. XPress allows conversion to a format called "Xpress markup". This format enables to spot the different structures of the book (if the Xpress file has been laid out well - which is not always the case). The structure of each book had to be worked out to create the perl script which enables the translation to the PAROLE format. Conformance to the PAROLE format was made thanks to a "nsgmls" tool. The errors found during the verification have been manually corrected.

For more information on prices for the PAROLE French Corpus, please contact ELRA.

CORRECTION

The prices for Dutch PAROLE corpus and lexicon (ELRA-W0019 and L0031) have changed.

Prices in euro	Corpus		Lexicon	
	Mb	NMb	Mb	NMb
R	270	300	300	400
RC	800	1,300	1,600	3,000
C	1,600	2,500	8,000	10,000

R: for research use by academics
RC: for research use by a commercial organisation
C: for commercial use
Mb: ELRA members
NMb: non members

For academic users from the Netherlands and Belgium, please contact ELRA.

LAST MINUTE ANNOUNCEMENTS

Now available:

- Danish SpeechDat(II) FDB-1000 (ELRA-S0072) and FDB-4000 (ELRA-S0073) databases.
- Swedish SpeechDat(II) FDB-1000 (ELRA-S0071) and MDB-1000 (ELRA-S0071) databases.

Available soon:

- PAROLE Italian corpus and lexicon.

For more information, please contact ELRA or visit the ELRA Web site (http://www.icp.grenet.fr/ELRA)