

Looking Forward to FP6

Steven Krauwer, ELSNET Coordinator, Utrecht University



Members of the ELSNET Board enjoying a meal together. Pictured (left to right) are Niels Ole Bernsen, Joseph Mariani, and Steven Krauwer

Even though we still have some two years to go under the current Fifth Framework Programme (FP5) of the European Commission, preparations for its successor, FP6, are already well advanced. I would like to mention three characteristics of the new programme that seem to be highly relevant for our community:

- FP6 projects will generally be fewer in number and much larger in size
- language and speech technology will not be part of the programme as distinct topics, but will appear as integrated components of broader themes, notably (but not exclusively) information management and interfaces
- even if the term “Network of Excellence” is maintained, it will refer to a totally different concept. Instead of comprising broad communities of key players in the field, these networks will be restricted to smaller consortia of top players in specific areas in which Europe has a leading position, carrying out research programmes aimed at strengthening Europe’s position in these areas

How will this affect ELSNET? It is a bit early to answer this question precisely. Experience has taught us that every single framework programme comes with its own thematic priorities, working structures, and funding conditions, and it is too early to see the full picture, let alone the precise position of our community within it. Yet it is clear that there are things we can do in order to prepare ourselves for the future, and there is still time for us to do that, as the present ELSNET funding contract has been extended to July 2002, and the successor contract – the last one under FP5 – will start on August 1 this year, and end on February 1, 2004.

In order to give us a better starting position for FP6, we have made a number of adaptations in our organisation and in our work programme for the next two years.

First of all, we have strengthened the present ELSNET Executive Board by including three new members from industry: Jean-Pierre Chanod (XEROX, Grenoble); Tony Rose (Reuters, London); and Denis Johnston (BT Adastral Park, Ipswich). Hans Uszkoreit (DFKI, Saarbrücken) has joined the Board as a new academic member. >

CORETEX
Marcello Federico
Michael Pitz 3

**The Ferret: Plagiarism
Detection System**
Caroline Lyon 5

ELSNews Inter views
Walter Daelemans 6

**Jewels Website
Announcemet**
Gerrit Bloothoof 8

Opinion Column
Kenneth Church 9

**Update on Activities
of CNLP, New York**
Elizabeth D. Liddy 10

**Announcements:
ELSNET Roadmap
Workshops;
10th ELSNET
Summer School** 12

**SIGdial Page:
DARPA
Communicator**
Alexander Radnicky 13

**NATO Advanced
Study Institute
Announcement** 14

Future Events 15

**Winter
2001-2**

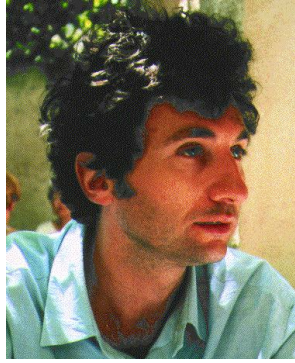
elsnet
.....

Secondly, we will focus on a few activities that seem to be of strategic importance for the future. Our priorities will be training and roadmapping.

Training has always been one of our strengths, and it is unclear to what extent training activities will be specifically supported by the FP6 funding schemes. It does seem very clear that the new, embedded position of language and speech technology in FP6 calls for a new approach to training, with a strong emphasis on the application and applicability of language and speech technology in specific contexts rather than seeing them as stand-alone activities. The choice of the topics for our next two summer schools reflects this change: in 2002 the topic "Evaluation and Assessment of Text and Speech Systems" is aimed especially at integrated language and speech systems in various application contexts [see announcement of this event on page 12 of this issue – Ed]; in 2003, "Language and Speech Technology for the Web" will target developers of web applications and services.

Our roadmap activities are already well under way: so far we have run workshops in Katwijk, Toulouse, (ACL/EACL 2001, Roadmapping for Human Language Technologies in Knowledge Management), and Santiago de Compostela, (MT Summit 2001, Roadmapping for Machine Translation). Reports of these events will be published shortly on the ELSNET website. Other roadmap workshops are being organised: for example, in March 2002 there will be another one on MT (in conjunction with TMI 2002 in Keihanna, Japan); and in June 2002 we are running a Language Resources Roadmap workshop in conjunction with LREC 2002 in Las Palmas (Canary Islands) [see announcements of these two events on page 12 of this issue – Ed].

Over the course of 2002 we will set up an integrated, interactive website which will both serve to publish our results (in documents and graphical representations) and to elicit feedback from the community. The technological roadmaps, which will be gradually integrated and extended, are expected to provide not only a common vision of the main challenges ahead and ways of meeting these challenges, but also a useful instrument in identifying European research priorities and strengths when creating potential FP6 project consortia and networks of excellence.



Other ELSNET activities will be our continued involvement in resources and evaluation (also in connection with our training and roadmapping actions, as above), with special emphasis on international collaboration, and information dissemination within the community via our website, email lists, and *ELSNNews*.

Let me finish with an invitation to all the members of our community to participate actively in our activities. Much of what we do is crucially dependent on your cooperation: we can't run a summer school without teachers, organisers, or students; we can't create a broadly-supported roadmap if you don't send us your contributions or your feedback; and we can't make information accessible to others if you don't communicate it to us in the first place.

ELSNNews welcomes the new members of the ELSNET Board.

Pictured left, from top to bottom:

Jean-Pierre Chanod (Executive Board)

Tony Rose (Executive Board)

Denis Johnston (Executive Board)

Hans Uszkoreit (Academic Board)

FOR INFORMATION

Steven Krauwer is Coordinator of ELSNET, at Utrecht University

Email: steven.krauwer@let.uu.nl

ELSNET web site:
<http://www.elsnet.org>

For advanced notice of the two roadmapping events mentioned and the ELSNET Summer School, see page 12 of this issue

ELSNNews is published at the School of Cognitive and Computing Sciences, University of Sussex. It is printed by the University of Sussex Print Unit.

Editor: Jenny Norris.

Editorial Team: Geoffrey Sampson, Steven Krauwer, and Brigitte Burger.

ISSN 1350-990X

© ELSNET 2001

FOR INFORMATION
Contributions to
ELSNNews, and corrections
should be sent to:
jennyn@cogs.susx.ac.uk

Tel: +44 1273 678659

Fax: +44 1273 671320

**Material for the next
issue is due:
18 March 2002**



*Marcello Federico, ITC-IRST, Trento, Italy, and
Michael Pitz, RWTH, Aachen, Germany*

The European Project CORETEX

Feature

CORETEX is the name of a European project funded within the IST programme running under the Fifth Framework. The project aims at improving core speech recognition technologies, which are central to the most important applications involving voice technology, such as multimedia information access and automatic services run over the telephone network.

The project consortium consists of four world-class speech recognition laboratories, namely: RWTH (Aachen, Germany); UCAM-DENG (Cambridge, UK); ITC-IRST (Trento, Italy); and CNRS-LIMSI (Paris). CORETEX officially started in April 2000 and will run until March 2003. The project has also set up a user group which includes representatives of European companies operating in the fields of information technology, telecommunication, broadcasting, and multimedia archives.

Objectives

The last decade has seen impressive advances in the capability and performance of speech recognition technology. Today, state-of-the-art systems around the world can feature speaker-independent, continuous speech recognition, capable of transcribing the audio of broadcast news with error rates below 20%. These advances can be explained by the increased accuracy and complexity of the statistical models employed, which are related to the availability of large speech and text corpora, and by the development of better training and decoding algorithms, which can today rely on much faster and cheaper computers.

Despite the extent of technical progress and the availability of commercial speech recognisers, many remaining problems still need to be addressed before the speech recognition problem can be considered "solved". For instance, speech recognisers are fairly sensitive to the acoustic and linguistic properties of the data, and in particular to mismatches between the training and the real usage conditions. Moreover, porting to a new language or task still requires substantial investment of time and money, and requires the knowledge of speech recognition experts.

CORETEX is addressing three relevant and still open problems related to speech recognition technology:

- **Genericness and adaptability** – the capability of speech recognisers to work properly on a wide range of tasks and to dynamically adapt using contemporary data

- **Portability** – the ability to port the technology to different languages and tasks at reasonable cost
- **Enriched transcription** – the provision of additional information to a simple textual transcription, which can be used as meta-data for indexing and retrieval purposes

Midway Results

Most of the research carried out during the first half of the project has been devoted to evaluating and improving the genericity, adaptability, and portability of speech recognisers.

Generally, for most of today's speech recognition systems, the amount of acoustic training material is crucial to its performance. Unfortunately, the generation of acoustic training material usually requires a lot of manual work and thus generates a great deal of the cost of developing a speech recognition system for new tasks or languages. The participating labs have investigated this problem from different perspectives and using different application domains and languages.

One possible way to reduce the work load is to manually transcribe only a few hours of the new acoustic material, to build a recogniser with these data, and to use this recogniser on larger quantities of un-transcribed training data. Depending on the level of supervision, recognition errors may occur in the automatically derived transcriptions. In the case of very low supervision, confidence measures help to improve the performance of the system trained on these automatically derived transcriptions. In the CORETEX project we have successfully developed methods that allow us to use only a small amount (one hour or less) of manually transcribed training data, which can be used to automatically derive transcriptions for un-annotated audio data.

The genericity of wide-domain acoustic models has been demonstrated by cross-task recognition experiments. Experiments have been carried out in American English to compare task-specific and cross-task models for the recognition of various types of data, including broadcast news transcription, a small vocabulary recognition task (TI-digits), dictation task (WSJ), and interactive dialogue task (ATIS). Further experiments have been performed on porting the Italian broadcast news system to different conversational

Winter
2001-2

elsnet



speech domains. In particular, the effects of adapting the acoustic and language models have been investigated – starting from either a small amount of manually annotated training data (up to two hours), or from a larger amount of recordings (eight hours), for which no manual transcriptions were available. Methods to improve genericity are under investigation, and these include multi-style training and unsupervised acoustic and language model adaptation.

Many important speech recognition tasks – in particular, spoken document retrieval – feature an open, constantly changing vocabulary. In a broadcast news task, for example, many unrecognised names of people and places occur. To address this problem, techniques have been developed which facilitate the automatic updating of the lexicon and the language model of the transcription system by exploiting newswire texts available on the Internet.

Recognition of (new) words requires acoustic base forms for them to be known. Commonly, words are transcribed manually, which constitutes a major burden on inter-domain portability. When orthographic transcriptions are available, grapheme-to-phoneme rules may then be applied. However, new words often exhibit irregular or foreign pronunciation. In the last year we have begun to develop methods for automatic transcription using morphological decomposition and statistical grapheme-to-phoneme conversion.

Future Work

Future work will continue the investigation on genericity, portability, and adaptability, along with three new topics: the extraction of information from audio streams; its representation in terms of metadata; and the establishment of an evaluation framework to perform cross-site comparison.

In general, the output of a speech recogniser is the best word string, an N-best list, or a word graph. However, other types of information can be extracted from the audio data and used to provide meta-data which can be used for further processing. The CORETEX project will investigate techniques for producing a rich symbolic transcription of the audio input, aimed at providing extra information for higher level processing. This meta-data will provide information about the acoustic source and recording environment (e.g., telephone or wide band, quiet or noisy), the type of signal (speech/music/background), speaker characteristics (gender, identity, dialect), and linguistic content. The linguistic content information will include cased text output, some forms of punctuation, speaker turns, and named entities or other semantically meaningful units.

Work has already started on the automatic detection of names of people, organisations, and locations occurring in Italian and English broadcast news audio archives. In

particular, we are focusing on techniques which allow us to bootstrap the recognition of named entities from little-supervised data and to reliably detect novel occurrences of entities from contemporary news texts. Related research topics that will be investigated in the future are story segmentation and topic spotting. Other efforts will be directed at extracting speaker information from large audio archives for the purposes of indexing, retrieval, and data mining.

Evaluation, of course, plays a crucial role in assessing progress. The partners in CORETEX are well aware of the importance of careful experimentation and well-defined evaluation. They regularly participate in organised evaluation benchmarks (DARPA, CLEF, SqaLe, and Aupelf), which will be an essential aid in defining the evaluation methodology and experimental set-up adopted by the project. For each research topic, objective evaluation criteria and measures need to be defined, test suites must be collected, and, if required, software tools will be developed. The differences between the four speech recognition platforms made available by the consortium is to be exploited in order to provide insights into the performance of the core technology. In particular, evaluation will investigate the relevance of different acoustic, lexical, and language model settings in the development of a new system, which may be developed from scratch or using existing components.

Evaluation metrics for enhanced transcription methods are also to be defined, in relation to indexing and information retrieval/extraction applications. Tools and benchmarks are to be developed to cover the languages of interest, and these will be in accordance with international *de facto* standards and in coordination with other international activities.

Finally, two showcases are to be developed during the coming year. The first of these is a special browser which will facilitate the examination of automatic transcripts of broadcast news with respect to both word error rate (by aligning manual and automatic transcripts at the word level) and metadata (by providing audio, speaker, and content information). The second showcase will apply the technology developed in CORETEX to the transcription of spontaneous speech conversations, such as meetings, interviews, lectures, talk shows, and so on.

FOR INFORMATION

Marcello Federico is a Senior Researcher at ITC-IRST, the Istituto Trentino di Cultura.

Email: federico@itc.it

Michael Pitz is a Researcher at RWTH Aachen, the Aachen University of Technology.

Email: pitz@informatik.rwth-aachen.de

CORETEX web site: <http://coretex.itc.it>



The Ferret – Ferreting out passages of similar text in large document collections

Caroline Lyon, University of Hertfordshire, UK



Caroline Lyon

It is surprisingly easy to fingerprint documents, so that each has a signature unlike any other. Passages of similar text in two documents can be found when fingerprints are compared and found to resemble each other.

Independently-written texts have different fingerprints, even when the same author is writing on the same subject on different occasions. This fact has been utilised in the development of a Plagiarism Detection system (commonly known as the Ferret) to find passages of similar text in students' essays. With hundreds of students each producing essays of several thousand words, it has become impossible to detect most instances of copying or collusion without using an automated system.

Using the Ferret a single similar paragraph can be found in any two essays (it does not have to be an exact copy). There are two steps to using the system: first, every document is compared with every other, and any that are suspiciously similar are flagged; then the user can select any pair of documents and display them side by side, with similar passages highlighted.

The system is in use at the University of Hertfordshire, where it has been developed by Caroline Lyon, James Malcolm, and Bob Dickerson, in the Department of Computer Science. It also works on other European languages, and has just gone on trial at Maastricht University. The initial verdict is that "The program did exactly what we expected and hoped for. it was 'a piece of (English) cake'. Very good."

Principles of fingerprinting text

The fingerprint of a piece of text is based on a large number of small, easily extracted features: word

trigrams, or overlapping three-word sequences. Each text is converted into the set of trigrams of which it is composed. When we consider separate texts on the same subject there will be certain common words, bigrams and trigrams (see Figure 1). However the

Classrooms have become inoculation centres as health workers try to stop the spread of the disease. More than **1700 pupils and staff** were injected today to combat what's been described as a **public health emergency**.

(35 words, 33 trigrams)

This morning children were queuing for injections not lessons at the school at the centre of the outbreak. Health teams have begun immunising **1700 pupils and staff** in an attempt to bring this **public health emergency** under control.

(38 words, 36 trigrams, 3 matches)

Figure 1: Example of two independent reports of the same news item. Matching trigrams are in bold. Note that most of the trigrams do not match.

phenomenon we exploit is the fact that common trigrams constitute a very small proportion of all the trigrams derived from independent texts. If the number of matching trigrams is above a certain threshold, then copying or collusion is suspected.

The first step in running the system can be done very quickly: using a standard desk-top PC, the processing of 120 reports totalling half a million words takes less than 20 seconds. The second step, when suspicious texts are displayed and compared, takes as long as the user wants.

(continued on page 7)

FOR INFORMATION

Caroline Lyon is a Senior Lecturer in the Computer Science Department at the University of Hertfordshire

Email: C.M.Lyon@herts.ac.uk

Web: <http://homepages.feisherts.ac.uk/~comrcml>

Feature

Winter
2001-2

elsnet
.....

Learning the Lazy Way

*ELSN*ews interviews *Walter Daelemans*

Walter Daelemans, director of the Centre for Dutch Speech and Language (CNTS) at the University of Antwerp, is one of the best-known exponents of the memory-based learning approach to language and speech processing. Daelemans and his team were responsible for the TiMBL suite of memory-based learning algorithms and metrics, which is distributed free for research use. Until 1999 Daeleman was at Tilburg University (Netherlands); he is now at Antwerp where he heads a 15-strong research team with his colleague Steven Gillis, a child-language expert, though he still spends one day a week at Tilburg where his former student Antal van den Bosch is taking the research forward.



Walter Daelemans

*ELSN*ews interviewed Daelemans about his work.

ELSNews: What led you to explore the memory-based approach?

Daelemans: In the mid-1980s I was working on speech synthesis; my background was in AI and knowledge-based systems. At the time, linguists mostly assumed that first-language acquisition involved inferring general rules from linguistic data (the rules had exceptions, of course). The Parallel Distributed Processing or “connectionist” crowd were starting to argue for a sharply different model, involving activation spreading through networks of neuron-like elements.

Thinking about text-to-speech conversion, I found both approaches unsatisfying. Rule-based models turned out to need so many exceptions, exceptions to exceptions, and so on, that they lacked all coherence. The connectionist experiments were more congenial, but they postulated mechanisms, for instance back-propagation, that were psychologically implausible and computationally slow. I wanted to find a way of modelling the material which was simple enough to be cognitively plausible, and simple enough to be practical.

ELSNews: How would you describe the essence of memory-based learning?

Daelemans: The core idea is that we do not use data of experience in order to infer general rules. What we store for the long term are the individual data items themselves. The approach is also called **lazy learning**: instead of immediately extracting abstract principles from our inputs, we just store the inputs until we need to use them for processing a new case. Then we handle the new case by analogy with the past examples in terms of a “*k* nearest neighbours” pattern-recognition algorithm.

We extended the same approach to a symbolic domain.

In any realistic research scenario involving speech, rule-based approaches are problematic not just because rules have exceptions, but because genuine data are contaminated by “noise”. Rather than trying to purge noise from one’s model, it is best to retain everything, noise included. We have shown again and again that, in language processing, “forgetting exceptions” damages system performance.

Nowadays, of course, we are not the only group working along similar lines. Royal Skousen in the USA was a pioneer in analysing language behaviour in terms of analogies rather than rules. Nagao’s example-based machine translation technique has a clear affinity to memory-based learning. Closer to home, work by Remko Scha and Rens Bod on Data-Oriented Parsing shows a strong family resemblance to what we are doing, and Harald Baayen at Nijmegen has been using our TiMBL software.

ELSNews: When you describe memory-based learning as a practical model, does that mean that it has industrial applications?

Daelemans: Yes it does. Our collaborator Jakub Zavrel is using this work as the basis for his spin-off company **Textkernel**, which deals in information extraction systems applied to material such as job-vacancy adverts or CVs. Jakub began in 2000, and his company is still with us (which is more than can be said for a lot of IT-based startups, as we approach 2002!) We have been selling licences for the commercial use of language-engineering components such as taggers, including sense-taggers, and morphological analysers, and for memory-based learning applications in other areas.

By “practical” I was not referring only to industrial exploitation, though. Our approach is also giving good results in well-defined, quantitative language processing

e

tasks, even though industrial exploitation may be some way off yet. For instance, our word-sense disambiguation system achieved good scores in the two SENSEVAL disambiguation competitions held to date, and – what is really more significant than the high placing we achieved – those good results required relatively little special-purpose adaptation of the general learning software.

ELSN*ews*: You say that your model of language behaviour is psychologically plausible, too. Can you illustrate that?

Daelemans: Take the case of assigning word stress in Dutch. We tried to model speakers' behaviour with a rule-based system, and it didn't work. A memory-based learning system worked well in general, and in particular it succeeded in modelling a rather surprising aspect of behaviour. Dutch speakers sometimes regularise stress patterns that are irregular, but also, they sometimes "irregularise" regular patterns. That second thing is very hard to explain in a rule-based framework.

ELSN*ews*: What next?

Daelemans: We have many plans: one exciting prospect relates to prediction in place of retrodiction. Our models have been successful at matching observed speaker behaviour, but to date this has been behaviour which we already knew about before we built the models. In collaboration with the psycholinguist Dominiek Sandra's group, we are beginning to develop models that will yield predictions about behaviour that we haven't yet studied. We shall observe the behaviour only after we

have the predictions. If they then turn out to match, this should be a specially persuasive vindication of our approach.

ELSN*ews*: It should indeed!



Some members of the CNTS group

Back row: Guy De Ruw, Gert Durieux, Veronique Hoste, Masja Kempen, Walter Daelemans

Front row: Anne Kool, Hanne Kloots, Helena Taelman, Steven Gillis

FOR INFORMATION

Email: walter.daelemans@uia.ua.ac.be

Web: cnts.uia.ac.be/~walter/

CNTS home page: cnts.uia.ac.be

TiMBL demos: ilk.kub.nl

Textkemel: www.textkemel.nl

(continued from page 5)

Spin-off from speech recognition applications

The Ferret is a spin-off from work at the University of Hertfordshire on the development of real-time subtitles for live TV programmes – now in use for sports programmes on BBC2. A professional speaker dictates the subtitles into a speech recognition system, which turns out the subtitles. To achieve the very high recognition rates that are necessary, customised language models have to be developed for each sport, and these language models are based on word trigrams.

A fundamental issue in the development of language models is the sparse-data problem: the well documented fact that a few trigrams are common, but most are rare. For example, a corpus of 39 million words from the *Wall*

Street Journal was examined. In this limited domain, with articles written in a prescribed style, the percentage of trigrams in a new article that had not appeared in the base corpus was, on average, 77%.

Now this sparse data problem can be stood on its head and utilised to fingerprint text – which is precisely what we have done to produce the Ferret.

Continuing work

We are extending the use of the Ferret to incorporate analyses of web pages, and we plan to move beyond the Higher Education arena.

**Winter
2001-2**

elsnet
.....

Jewels Website Launched

Gerrit Bloothoof, Utrecht

The Joint European Website for Education in Language and Speech is now accessible at www.hltcentral.org/jewels. Although work is still continuing on the site, it is worth paying a visit. There are seven main sections:

General

Here a visitor can learn about language and speech. The current presentation is geared at people already working in the field, and gives an overview of areas in language and speech, their elements of study, suggested reading, web links, and related products. Work is under way on presentations aimed at the interested public and prospective students.

News

This section presents the latest news, an agenda of events such as summer schools, and an archive of past events.

Information

If you want to know about European programmes for Higher Education, this is the spot. There is also advice on student exchange and mentoring schemes.

Recommendations

Following on from the Socrates Thematic Network projects, the site gives the recommendations that have been made for education in the areas of phonetics, spoken language engineering, computational linguistics, and speech and language therapy. The recommendations are presented for several audiences, from institutions to EU bodies.

Institutions & Courses

This section is still being developed, but the aim is to give a comprehensive presentation of all institutions in Europe that provide education in language and speech.

Ideally – and for this we will need readers' help – there will also be links to all courses that are running across Europe. There is a search function which facilitates searches for words at one level deep in institutional web pages.

Tools & Materials

What tools and materials are available to support a particular need? JEWELS has started to create a database with descriptions of, and links to, resources that should help you to find the support you need. Comments from experienced users will be available to help you to make the best choices.

Search

Search through the whole JEWELS site and its databases, but also at the level of linked web pages (one level deep).

An editorial team is responsible for updating JEWELS, but will need your help to enhance the contents and to keep the information up-to-date and of high quality.

JEWELS is the result of the Socrates Thematic Network projects "Speech Communication Sciences" and "Advanced Computing in the Humanities", and is currently supported by ELSNET. The website has been built by VDI-VDE in Berlin.

FOR INFORMATION

For more information about JEWELS contact
Gerrit Bloothoof

Email: gerrit.bloothoof@let.uu.nl

JEWELS web site: www.hltcentral.org/jewels

(continued from page 9)

into a glorified typewriter? And the ultimate indignity is that "the marketing genius" that gave us this "new and improved" typewriter has become the richest man in the world! Recall, at the time, no one ever got fired for buying Big Blue, and ballistics was a killer application.

Typewriters are pretty neat things. My four-year-old son has known computers all his life, but he just recently inherited a typewriter from his grandmother. To him, the given/new relationship is reversed: computers are old hat, but his "new" typewriter is a technological marvel. You can actually see how it works! Now, that's a user interface worth telling mummy about.

By comparison with word processing, search has even more potential. It is becoming increasingly difficult to separate one part of the web (e.g. desktop publishing) from another (e.g. search). The real service to society is getting people to communicate with one another, whether that be by improving the technology for reading or the technology for writing (or for talking or for listening). In short, I think we can all feel good about having made a real contribution to the arms race that matters. Microsoft and other major companies have been investing heavily in our field for a long time because they know there's a killing to be made in communications.



These are Hard Times

Opinion Column

Kenneth Church, AT&T Labs Research, USA



In this issue we are delighted to welcome our new columnist, Dr. Ken Church, who will be giving us his thoughts over the next few months.

We extend our thanks to our out-going "Opinionist", John Nerbonne, from the Rijksuniversiteit Groningen in the Netherlands, who has contributed regularly to this column over the past year. Many thanks, John!

This is a huge change from just a few years ago, when everyone was investing like crazy (quite literally) in what we do.

The basic assumption behind the so-called (and perhaps soon to be forgotten) "new" economy is that demand is everything. If you get a lot of people to use your products or visit your website (even if you have to pay them to do so) then eventually you'll get rich. The problem with this logic is that you don't want to end up like Crazy Eddie, a retailer in the New York area who advertised that his prices were insane – and apparently they were.

Whilst in retrospect it might seem absurd to sell at below cost (with negative margins), the idea is not completely without merit. The newspaper industry has long recognised that circulation is key to what they do. So too, should we proudly acknowledge that our having a positive (for the most part) impact on the everyday lives of billions of people around the world has got to be good for something, at least in the long run.

At any rate, back when the ALPAC report was published (1966), the experts in the field could already see a few financially rewarding applications of what we do (Appendix 17). I will refer to these applications as word processing, desktop publishing, and information retrieval, though of course, it wasn't until some time later that this terminology became buzzword compliant. Of course, those ALPAC predictions came true (and more so). Contrary to the standard critique of the ALPAC report (see <http://ourworld.compuserve.com/homepages/WJHutchins/Alpac.htm>), its biggest mistake was not so much its position on machine translation (which had a big impact within the field, but relatively little elsewhere), but that it sandbagged the very applications that it was hawking. (How's that for a feel-good rational reconstruction of history!)

In retrospect, it seems obvious that Vannevar Bush got it right with *As We May Think* (1945) (<http://www.isg.sfu.ca/~duchier/misc/vbush/vbush.shtml>), but at the time of the ALPAC report, who would have thought the almighty mainframe (and the priesthood) would evolve

(continued on page 8)

New York City, and the world with it, is in mourning. The events of September 11th brought out many emotions: shock, horror, grief. I have been pleasantly surprised by all the positive energy: everyone wants to help out in whatever small way they can; friends and colleagues around the world have been calling and emailing just to make sure everyone is OK. It is all very touching.

People don't know what they can do to help out. I have been helping to restore telephone services. It may not be much, but it's something. The mayor has been urging us to shop. That seems like an odd suggestion. No one feels like shopping. But the mayor has a point. These are hard times: unemployment is up; every day we hear about another dot-com about to dot-crash; and even big companies are feeling the pain; rumours of layoffs are everywhere. Perhaps all of this would have happened without September 11th, but I can't remember a time when jobs were so hard to find in our field.

It is easy to get depressed. It would be so easy to indulge in self-criticism. It would almost make me feel better to write an essay along the lines of the 1966 ALPAC report (<http://www.nap.edu/books/ARC000005/html/>), but I won't. That report was such a downer that funding for machine translation in the USA was severely cut back for decades. It isn't that I object so much to the ALPAC report itself – the report makes a lot of excellent points in a way that still resonate with the powers that be in my company (the chair of the ALPAC committee was an influential AT&T Bell Laboratories executive) – but rather, this isn't the time for singing the blues. I feel we should all be out there with the mayor trying to feel good again.

The fact is that basic research in speech and language is making a big difference in everyone's daily life:

- Nearly everyone uses search engines on the web
- Many people use speech recognition over the phone
- Lots of people now use machine translation on the web
- Everyone uses spelling correction on the PC, and many use grammar checkers (for better or for worse).

FOR INFORMATION

Ken Church is Head of the Department of Data Mining at AT&T Labs Research in New Jersey

Email: kwc@research.att.com

Web: <http://www.research.att.com/~kwc>

Winter
2001-2

elsnet



A Breadth of NLP Applications

Elizabeth D. Liddy, Center for Natural Language Processing, New York

Introduction

The Center for Natural Language Processing (CNLP) was founded in September 1999 in the School of Information Studies, the “Original Information School”, at Syracuse University. CNLP’s mission is to advance the development of *human-like, language-understanding software capabilities* for government, commercial, and consumer applications. The Center conducts both basic and applied research, building on its recognised capabilities in Natural Language Processing. Its seventeen employees are a mix of doctoral students in information science or computer engineering, software engineers, linguistic analysts, and research engineers.

Opportunities

We are finding today to be a particularly opportune time for NLP, due to the confluence of a number of factors:

- Sufficient R & D in the field of NLP has been accomplished in past years to provide solid baseline NLP capabilities
- Computational resources have caught up to the level of requirements of complex NLP systems
- The bulk of textual information that forms the basis on which all organisations conduct their business is now in electronic format.

These factors are key to both research centres and commercial vendors, and they contribute to the open, welcoming reception that NLP applications are now receiving from both funders and customers. NLP is proving itself as a powerful enabling technology for a range of applications supported by CNLP’s technology, including:

- Document Retrieval
- Question-Answering
- Information Extraction
- Text Mining
- Automatic Metadata Generation
- Cross-Language Retrieval
- Document Summarisation.

Be-Bee, Inc.

To meet the many opportunities before us, CNLP recently licensed in the **metaMarker** technology of Be-Bee, Inc., that will enable us to provide a wide range of implementations by building on this solid commercial NLP technology. With the increasing request for sophisticated language-handling applications from both government and commercial funders, we needed to focus on just those aspects of the technology that are research-related. The licensing in of Be-Bee’s ‘metaMarker’ advanced commercial capabilities has speeded up our time to delivery by bootstrapping our language processing

modules with accurate interpretations of input text and providing scalable and reliable language processing.

While there are other commercial technologies available, CNLP chose Be-Bee’s ‘metaMarker’ because of the quality and depth of its language processing – they are a real NLP shop. This is what they do, and they do it well. Also compared to the other technologies we looked at, Be-Bee’s technology is much more flexible and enables us to specialise the output in ways that matter to us for various applications, three of which we will describe.

Automatic Metadata Generation

Together with Be-Bee Inc., we are working on an automatic metadata generation project under funding from the National Science Foundation’s Science, Maths, and Engineering Digital Library Programme. The goal of our project is to run our combined natural language processing on learning resources (lesson plans, classroom activities, etc.) in order to assign values to the twenty-three metadata attributes that have been accepted as the metadata standard for education. These metadata values are then matched in response to users’ queries, or browsed by users to familiarise themselves with a digital library’s resources. The goal here is to break the metadata generation bottleneck which human assignment of metadata has caused. We will accomplish this by improving the speed with which educational resources can be made available; increasing the number of educational resources which are available electronically; and providing users with improved access to a digital library via richer and more complete metadata values.

In a second NSF Digital Library Project, we will be extending the metadata information we extract to include the educational standard that the resource can be used to accomplish. By working with a master list of educational content standards, which the state standards map to, this new metadata attribute will enable teachers and administrators in any state to select those teaching materials which will assist their students in meeting the required standards of their state.

Automatic metadata generation is an area of numerous opportunities because metadata standards are being developed and agreed in many domains, including business, geography, education, and biology, and for use by various technologies, including statistical-table browsers, digital libraries, and peer-to-peer technologies. If there is text associated with an object, NLP can provide the means of interpreting the text in order to understand whether the information for each metadata element is present in the object and then extract from the object the values to fill the appropriate standard’s metadata record. This uniform description will then make the object findable and accessible by users.



Question-Answering

Question-Answering (QA) is currently a very hot topic, and has a breadth of appeal because it is an application with many flavours based on the context in which the QA system is used.

Building on our existing eQuery system, we are currently developing a QA capability for NASA for use within a collaborative learning environment for distance education in aeronautical and mechanical engineering. The students will use the collaboration

technology for class interactions and for group project work. At any time, a student can pose a question, phrasing it as naturally as if they were asking the professor or another student. In turn, the QA system interprets the query at all the levels of language at which humans extract meaning.

For example, for the query "What is the best material for the wings of reusable spacecraft?" the Language-to-Logic module of our eQuery system would produce the following representation:

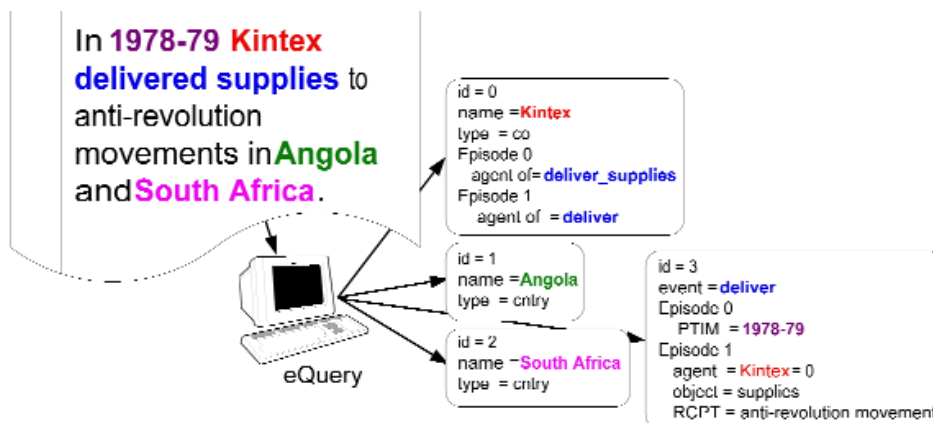
```
material* AND wing* AND reusable_spacecraft
OR reusable_launch_vehicle OR RLV
```

in which the correct logical relations are understood, conceptual phrases are recognised, and single words are stemmed to match morphological variants, as well as being expanded to include their synonymous phrasings. The query representation is searched against the NLP-indexed technical papers, class lectures, questions and answers accumulated over time (Previously Asked Questions / PAQs), and transcripts of problem-solving interactions of prior classes. eQuery then presents the student with one or more answer-providing passages, which the system has ranked according to their likelihood of containing the answer to their question.

Information Extraction

Government and competitive intelligence is another application area where we have utilised the rich interpretation provided by our NLP for Information Extraction (IE). We have found that both sets of customers need to extract a broad range of entities from text. Our IE includes the more typical person, place, and location slots, but also the more specialised extractions (e.g. perpetrator, drug, weapon, backer) – the specialisation depending on the domain of interest, for a total of 165 entities. Our technology also recognises and extracts the relations between or amongst entities, and most recently we have been focusing on using a frame-based representation to recognise and extract the multiple

aspects of events. The accompanying figure provides an example of the types of extraction our eQuery Document Processor performs. In our current Evidence Extraction and Link Discovery Project for DARPA, we



are focusing on scenario extraction, which requires learning the set of events which predicted an incident of interest in the past, and then recognising when this scenario appears to be playing itself out again.

A project that we recently completed for Unilever, a large international conglomerate, focused on adapting our IE technology for their use in competitive intelligence by enabling them to track their competitors' activities from web and subscription database sources. The IE capability processes large-volume daily news-feeds, recognising, interpreting, and extracting entities, relations, and events of interest and feeding them into a visualiser to make it possible for their strategic-intelligence staff to recognise patterns that may not be apparent when individual documents are processed.

Future Work

While the existing set of applications for NLP is broad, varied, and interesting our belief is that there will be an even larger number of viable applications in the near future. As speech-understanding technology improves and voice-input applications multiply, the need for full NLP capabilities will grow exponentially. When we think of it, a very high proportion of what humans accomplish is either accomplished through language or is reported in language – and therein lies the future of NLP applications.

FOR INFORMATION

Professor Liddy is Director of CNLP, School of Information Studies at Syracuse University in New York

Email: liddy@mailbox.syr.edu

Web: <http://www.cnlp.org>

Be-Bee, Inc. web site: www.be-bee.com

**Winter
2001-2**

elsnet

.....

ELSNET Roadmapping Workshops Coming Soon

Towards a Roadmap for Multimodal Language Resources and Evaluation

An ELSNET workshop at LREC 2002, Las Palmas, Canary Islands, Spain, on June 2, 2002

The workshop aims to bring together key players in the field of resources and evaluation to work towards creating a Roadmap for Language Resources – a broadly-supported view on the longer, medium, and shorter term needs and priorities. This lies within the context of ELSNET's other roadmapping activities, which aim at developing a technological roadmap for the whole field of Human Language.

The workshop should produce a draft roadmap report to be published on the web in order to elicit feedback and to serve as an example for the creation of other roadmaps. Sessions may include invited talks, a small number of refereed presentations, panel sessions, and small working groups addressing specific questions.

We are inviting position papers aimed at identifying or interconnecting key issues and resources, major challenges, and strategies to reach our goals, as well as papers addressing meta issues (e.g., what is a roadmap?, how to reach consensus, etc).

For submission details visit the workshop web site: www.elsnet.org/roadmap-lrec2002.html, or contact Steven Krauwer, ELSNET Coordinator: s.krauwer@elsnet.org.

Conference web site: <http://www.lrec-conf.org>

MT Roadmap Workshop

An ELSNET workshop on March 16, 2002, at TMI2002 (March 13-17), Keihanna (near Kyoto), Japan

The workshop aims to reach a broadly-supported definition of a context in which to position the MT community's efforts, to allow us to identify common priorities for joint activities in areas such as research, resources, and training. It is intended for people with an analytical or future-oriented, programmatic interest, from both research and industry.

The workshop will comprise three component sessions:

- composing a critical analysis of the present state of machine translation in the broadest sense
- setting visions of the future
- identifying major research challenges and establishing intermediate goals.

We invite papers that relate to any of the above topics.

Sessions may include invited speakers or panels, submitted papers (reviewed), and ample space for discussion.

The results will be published on the ELSNET web site, to form the basis of a broad consultation of the MT community on the future directions of machine translation in the broadest sense.

For submission details visit the workshop web site: www.elsnet.org/roadmap-tmi2002.html, or contact Steven Krauwer, ELSNET Coordinator: s.krauwer@elsnet.org.

Conference web site:
<http://www.kecl.ntt.co.jp/events/tmi/>

10th ELSNET European Summer School on Language and Speech Communication

This year, the ELSNET European Summer School on Language and Speech will be held from **July 15-26**, with the topic **'Evaluation and Assessment of Text and Speech Systems'**. The school will provide courses on evaluation of language, speech, and multimodal systems and components, and on the use of corpora and annotation schemes, methods and tools in the evaluation process. It is aimed at advanced undergraduates, PhD students, postdocs, and academic and industrial researchers and developers, and is organised by the Natural Interactive Systems Laboratory (NISLab), University of Southern Denmark, Odense.

Grants will be available from the EU Improving Human Potential programme and may cover a substantial part of total costs for young European researchers. The ISCA grant scheme is open for applications related to this event.

Course topics

- Evaluation of spoken dialogue systems
- Speech recognition evaluation: broadcast-news and meeting transcription, audio-visual speech recognition
- Evaluation of question/answering, summarisation, and information retrieval systems
- Evaluation of machine translation systems
- Annotation and evaluation of text and speech systems
- Evaluation of talking animated interface agents
- Resources and evaluation of text and speech systems
- Evaluation of part-of-speech taggers and/or parsers and/or terminology and/or lexicons, etc.
- Usability evaluation
- Evaluation of speech synthesis systems.

FOR INFORMATION

Web: <http://summerschool2002.nis.sdu.dk/>

Phone: (+45) 63 15 73 08

Fax: (+45) 63 15 72 24

Email: hensen@nis.sdu.dk

SIGdial: The DARPA Communicator Programme

Alexander I. Rudnicky, Carnegie Mellon University, USA



SIGdial
(ACL Special
Interest Group
on Discourse and
Dialogue)

In 1999, DARPA initiated a programme directed at the development of advanced spoken dialogue systems. At that time it was perceived that whilst advances had been made in the development of technologies for spoken language interaction (for example, in the ATIS programme) and different research systems had been developed with the purpose of enabling human dialogue, no systems were available that could be applied to realistic task scenarios. The Communicator programme therefore focused on the development of systems that could support “real-world” complex problem-solving activities, using travel as the target domain. All sites built systems that accessed actual travel information (including air, hotel, and rental cars), either by collecting it from the web or through subscription to a commercial service. All sites also sought to create services that would be of value to real users, as opposed to only being exercised by the developers or by hired testers. For example, users could register with a particular site to receive by e-mail the itineraries they had created.

A second goal of the programme was to develop a dialogue system architecture that could be used both by programme participants and by the community at large. The purpose of a common architecture was to enable interchange of component modules between sites and to allow sites that did not have the resources to construct entire dialogue systems (but were interested in working on particular components) that could field a complete system. This architecture, “Galaxy”, has provided a stable environment for participating sites (and others) and has been made available in Open Source distribution (see <http://communicator.sourceforge.net/>). A number of participating sites are also making their systems available, under a variety of licensing arrangements.

The programme undertook two major data collections, for evaluation purposes, one in the summer of 2000 (nine sites) and the other in the summer of 2001 (eight sites). In both cases, users were hired through the testing organisation (NIST) and instructed to call Communicator systems. Each system generated a standard log (a benefit of sharing the Galaxy architecture) that was comparable across sites and allowed the computation of uniform statistics. Users filled out standard web-based questionnaires after each of their calls. The Summer 2000 evaluation was organised using a within-subjects design. That is, a panel of testers called each system in turn and attempted to construct an itinerary using either a fixed scenario constructed by NIST or their own personal travel plans.

Analysis of the data from this evaluation revealed a number of shortcomings, principally the use of fixed scenarios (representing artificial travel, which users may have had trouble engaging with) and the requirement that a given user call multiple systems (leading to potential user confusion about system capabilities).

For the Summer 2001 evaluation, a marketing firm was engaged to hire a panel of frequent travellers who were instructed to call a particular system to arrange their own travel. In addition, they carried out fixed scenarios, to calibrate 2001 systems with respect to 2000 systems and to thoroughly probe system functionality.

Because each site adopted a somewhat different design, the evaluations provided a unique opportunity to understand what strategies contribute to effective dialogue. For example, the first evaluation revealed that systems adopting a directed dialogue strategy were generally more successful (as opposed to systems that supported a more mixed-initiative strategy). There appear to be two reasons for this: directed dialogue, with the concomitant constraint on the range of user input, makes the system less susceptible to recognition error (the most common source of task failure in Communicator). Secondly, directed systems provide task structure for the user, reminding them of the information needed for completing the task, and correspondingly relieving the user of memory load. As it turns out, this advantage is particularly evident in the constraint-gathering phases of the task (e.g., finding out the user’s destination) and is less useful in the solution navigation phase (e.g., choosing between alternative flights or re-stating constraints) when the user requires greater flexibility. Systems adopting a directed strategy for the latter phase generally exhibited relatively poorer task completion and less user satisfaction.

Once prepared for distribution, the corpus collected during the evaluation will be made available to the wider research community through the Linguistic Data Consortium (<http://www ldc.upenn.edu/>), as a resource for the study of human-computer interaction by spoken language.

FOR INFORMATION

Alexander Rudnicky is a Senior Systems Scientist in the School of Computer Science at Carnegie Mellon.

Email: air@cs.cmu.edu

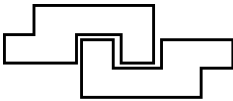
Web: www-2.cs.cmu.edu/afs/cs.cmu.edu/user/air/WWW/

SIGdial web site: <http://www.sigdial.org/>

**Winter
2001-2**

elsnet





Advanced Study Institute

Dynamics of Speech Production and Perception

June 23 – July 6 2002; Il Ciocco, Italy

Applications for participation in a NATO Advanced Study Institute (ASI) on the Dynamics of Speech Production and Perception are being solicited. The objective of this ASI is to expose young scientists from diverse disciplines to facts, theories, and issues related to speech dynamics. It will survey both classical and contemporary speech research, with dynamic processes as the common focus, from the following perspectives:

- articulatory phonetics and physiology
- auditory perception of speech and other dynamically changing signals
- neurophysiology of the central auditory system
- spoken language processing
- speech technology

The ASI will include lectures by established scientists, round-table discussions, demonstrations, and poster presentations by participants. The language of the ASI will be English.

Lecturers will include René Carré (France; articulatory and perceptual phonetics, speech acoustics, gesture theory), Pierre L. Divenyi (USA; psychoacoustics), Steven Greenberg (USA; spoken language processing, auditory physiology), Hynek Hermansky (Czech Republic and USA; speech recognition, signal processing), András Illényi (Hungary; acoustics), Björn Lindblom (Sweden; linguistics, speech physiology, speech perception), Valeriya V. Lublinskaya (Russian Federation; speech perception, speech physiology), Georg Meyer (U.K.; computational neuroscience, psychoacoustics), Israel Nelken (Israel; auditory physiology), Roy Patterson (U.K.; models of hearing, psychoacoustics), Louis Pols (The Netherlands; perceptual phonetics, speech acoustics), Elliot Saltzman (USA; articulatory phonetics, complex systems analysis), Shihab A. Shamma (USA; auditory physiology, computational modeling), Victor N. Sorokin (Russian Federation; control theory, acoustics, speech production), and Klára Vicsi (Hungary; psychoacoustics of speech, speech recognition).

Participants are expected to specialise in one or more areas related to speech, such as phonetics, speech technology, experimental psychology, physiology, and

acoustics. Individuals who have recently completed or are near to completing their studies at institutions of higher learning are encouraged to apply, although applications by scientists of more senior standing could also be considered under special circumstances. Although the ASI is primarily directed to citizens of NATO, Partner, and Mediterranean Dialogue countries, qualified scientists from other countries may also apply.

There is no registration fee. Selected participants who otherwise would be unable to attend, especially those from Partner countries, may apply for partial subsidy to defray their accommodation and travel costs. The number of participants is limited.

The “Centro Internazionale Il Ciocco”, situated in the Apennines north of Lucca (Tuscany), is an ideal venue for intensive workshops. It is a quiet mountain resort with first-rate conference facilities, comfortable rooms, excellent Italian food, as well as modern recreational and sports facilities.

Interested individuals should initially contact Klára Vicsi (see below) regarding space at the event and deadlines. Applications should then be sent by e-mail to asi2002@ebire.org. Applications should include:

- contact information
- a recent Curriculum Vitae
- a list of relevant publications
- a brief statement on how participating in the ASI would benefit the applicant’s educational and career objectives.

Organising Committee:

Pierre L. Divenyi, Director;
Klára Vicsi, Co-Director.

FOR INFORMATION

Centro Internazionale Il Ciocco web site:
<http://www.ciocco.it/>

Email: vicsi@tft-202.ttt.bme.hu

Future Events

- March 13-17** *9th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2002)*: Keihanna, Japan.
Email: teruko@cs.cmu.edu URL: <http://www.kecl.ntt.co.jp/events/tmi/>
- March 16** *Machine Translation Roadmap Workshop in association with TMI 2002*: Keihanna, Japan.
Email: s.krauwer@elsnet.org URL: www.elsnet.org/roadmap-tmi2002.html
- March 25-26** *Speech Technology in the Learning and Assistive Interface Symposium (InSTIL 2002)*: San Diego, California, USA.
Email: p.delcloque@msec.acuk URL: dbstaya.acuk/instil
- March 29-30** *The Asian-Pacific Workshop on Terminology (HKTerm 2002)*: Hong Kong.
Email: cthkterm@cityu.edu.hk URL: cpct92.cityu.edu.hk/TAHK/index.htm
- April 8-10** *ISCA International Tutorial and Research Workshop on Temporal Integration in the Perception of Speech (TIPS)*: Aix-en-Provence, France.
Email: tips@lpl.univ-aix.fr URL: www.lpl.univ-aix.fr/%7Etips
- April 11-13** *Speech Prosody 2002*: Aix-en-Provence, France.
Email: sp2002@lpl.univ-aix.fr URL: www.lpl.univ-aix.fr/sp2002/
- April 17-20** *The Processing of Arabic*: Tunis, Tunisia.
Email: brahem@irsit.rnrt.tn
- April 20-25** *Robustness in Speech Based Interfaces: Sharing the Tricks of the Trade. A workshop at CHI2001*: Minneapolis, Minnesota, USA.
Email: nilda@ida.liu.se URL: www.ida.liu.se/%7Enlplab/chi-ws-02/cfp.shtml
- April 24-25** *Language Technology for Business Information Systems – Special Session in conjunction with the 5th International Conference on Business Information Systems (BIS 2002)*: Poznan, Poland.
Email: piskorsk@dfki.de URL: biskie.ae.poznan.pl
- May 27-28** *Workshops in association with LREC 2002*: Las Palmas, Spain.
There are nine highly relevant pre-conference workshops covering many areas related to language resources and evaluation. For details, see “workshops” link on main conference web site.
URL: www.lrec-conf.org/lrec2002/
- May 29-31** *3rd International Conference on Language Resources and Evaluation (LREC 2002)*: Las Palmas, Spain.
Email: lrec@ilc.pi.cnr.it URL: www.lrec-conf.org/lrec2002/
- June 1-2** *Workshops in association with LREC 2002*: Las Palmas, Spain.
There are ten highly relevant post-conference workshops covering many areas related to language resources and evaluation. For details, see “workshops” link on main conference web site.
URL: www.lrec-conf.org/lrec2002/
- June 10-12** *The International Semantic Web Conference*: Sardinia, Italy.
Email: missikof@iasi.rm.cnr.it URL: iswc.semanticweb.org/
- June 17-21** *ISCA Tutorial and Research Workshop (ITRW) on Multimodal Dialogue in Mobile Environments (IDS'02)*: Kloster Irsee, Germany.
Email: laila@nis.sdu.dk URL: www.sigmedia.org/ids02
- June 28-29** *International CLASS Workshop on Natural, Intelligent, and Effective Interaction in Multimodal Dialogue Systems*: Copenhagen, Denmark.
Email: kuppevelt@ims.uni-stuttgart.de URL: www.dass-tech.org/events/NMI_workshop2.html

This is only a selection of events – see <http://www.elsnet.org/cgi-bin/elsnet/events.pl> for details of more events.

Winter
2001-2

elsnet



ELSNET

Office

Steven Krauwer,
Co-ordinator
Brigitte Burger,
Assistant Co-ordinator
Monique Hanrath,
Secretary
Utrecht University (NL)

Task Groups

Training & Mobility
Gerrit Bloothoof,
Utrecht University (NL)
Koenraad de Smedt,
University of Bergen (NO)

Linguistic & Speech Resources

Antonio Zampolli,
Istituto di Linguistica
Computazionale (I) and
Ulrich Heid, Stuttgart
University (D)

Research

Niels Ole Bernsen, NIS
Odense University (DK)
and Joseph Mariani,
LIMSI-CNRS (F)

Executive Board

Steven Krauwer,
Utrecht University (NL)
Niels Ole Bernsen, NIS,
Odense University (DK)
Jean-Pierre Chanod,
XEROX (F)
Björn Granström,
Royal Institute of
Technology (S)
Nikos Fakotakis,
University of Patras (EL)
Ulrich Heid,
Stuttgart University (D)
Denis Johnston, BT
Adastral Park (UK)
Joseph Mariani,
LIMSI-CNRS (F)
José M. Pardo,
Polytechnic University of
Madrid (E)
Tony Rose, Reuters (UK)
Geoffrey Sampson,
University of Sussex (UK)
Antonio Zampolli,
University of Pisa (I)

ELSNET P participants

Academic Sites

A	Austrian Research Institute for Artificial Intelligence (ÖFAI)
A	Vienna University of Technology
A	University of Vienna
B	University of Antwerp -UIA
B	Katholieke Universiteit Leuven
BG	Bulg Acad.Sci. - Institute of Mathematics and Informatics
BY	Belorussian Academy of Sciences
CH	SUPSI University of Applied Sciences
CH	University of Geneva
CZ	Charles University
D	Universität Stuttgart-IMS
D	Christian-Albrechts University, Kiel
D	Universität Hamburg
D	Institut für Angewandte Informationsforschung
D	German Research Center for Artificial Intelligence (DFKI)
D	Universitaet des Saarlandes
D	Universität Erlang-en-Nürnberg -FORWISS
D	Ruhr-Universitaet Bochum
D	Universität des Saarlandes CS-AI
DK	Center for Sprogteknologi
DK	University of Southern Denmark
DK	Aalborg University
E	University of Granada
E	Universitat Autònoma de Barcelona
E	Universidad Nacional de Educación a Distancia (UNED)
E	Universidad Politécnica Madrid
E	Universidad Politécnica de Catalonia
E	Universidad Politécnica de Valencia
EL	Institute for Language & Speech Processing (ILSP), Athens
EL	National Centre for Scientific Research (NCSR) "Demokritos", Athens
EL	University of Patras
F	LIMSI-CNRS, Orsay
F	LORIA, Nancy
F	Université Paul Sabatier (Toulouse III)
F	Université de Provence
F	Inst. National Polytechnique de Grenoble
F	IRISA/ENSSAT, Lannion
GE	Tbilisi State University, Centre on Language, Logic and Speech
HU	Lóránd Eötvös University
HU	Technical University of Budapest
I	Consiglio Nazionale delle Ricerche
I	Università degli Studi di Pisa
I	IRST, Trento
I	Consorzio Pisa Ricerche
I	Fondazione Ugo Bordon
IRL	Trinity College, University of Dublin
IRL	University College Dublin
LT	Inst. of Mathematics & Informatics
NL	Utrecht University
NL	University of Amsterdam (UvA)
NL	University of Nijmegen
NL	Foundation for Speech Technology
NL	Tilburg University
NL	University of Twente
NL	Eindhoven University of Technology
NL	Universiteit Leiden
NL	Netherlands Organization for Applied Scientific Research TNO
NL	University of Groningen
NO	University of Bergen
NO	Norwegian University of Science and Technology
P	University of Lisbon
P	New University of Lisbon
P	INESC ID, Lisbon
PL	Polish Academy of Sciences
RO	Romanian Academy
RU	Russian Academy of Sciences, Moscow
S	KTH (Royal Institute of Technology)
S	Linköping University
UA	IRTC UNESCO/IIP
UK	University of Cambridge
UK	Leeds University
UK	University of Sunderland
UK	University of Cambridge
UK	University of Sheffield
UK	University of York
UK	UMIST Manchester
UK	University of Essex
UK	University of Edinburgh
UK	University of Sussex
UK	University of Brighton
UK	University College London
UK	University of Ulster
UK	University of Dundee

Industrial Sites

B	Lemout & Hauspie Speech Products
D	Philips Research Laboratories

D	IBM Deutschland
D	DaimlerChryslerAG
D	aspect Gesellschaft für Mensch-Maschine Kommunikation mbH
D	Acolada GmbH
D	Grundig Professional Electronics GmbH
D	Novotech GmbH
D	Verlag Moritz Diesterweg GmbH
D	Sympalog Speech Technologies AG
D	Varetis Communications
D	Langenscheidt KG
DK	Tele Danmark
E	Sema Group sae
E	Telefonica I & D
EL	KNO WLEGDE S.A.
F	Memodata
F	Xerox Research Centre Europe
F	LINGA s.a.r.l.
F	Systran SA
F	LexiQuest
F	VECSYS
F	SCIPER
F	Aerospatiale
F	TGID
FIN	Kielikone Oy
FIN	Nokia Research Center
HU	MorphoLogic Ltd.
I	LOQUENDO
I	OLIVETTI RICERCA SCpA
LV	TILDE
NL	Compuer
NL	Knowledge Concepts BV
NL	Sopheon NV
NL	IP Globalnet Nederland BV
RU	ANALIT Ltd
RU	Russicon Company
S	Sema Infodata
S	Telia Pionator AB
UK	Imagination Technologies plc
UK	Canon Research Centre Europe Ltd
UK	Sharp Laboratories of Europe Ltd
UK	BT Adastral Park
UK	ALPNET UK Limited
UK	20/20 Speech Ltd
UK	Reuters Ltd
UK	Vocalis, Ltd.
UK	Hewlett-Packard Laboratories
UK	Logica Cambridge Ltd.
UK	SRI International

What is ELSNET?

ELSNET is the European Network of Excellence in Human Language Technologies. ELSNET is sponsored by the Human Language Technologies programme of the European Commission; its main objective is to foster the human language technologies on a broad front, creating a platform which bridges the gap between the natural language and speech communities, and the gap between academia and industry.

ELSNET operates in an international context across discipline boundaries, and deals with all aspects of human communication research which have a link with language and speech. Members include public and private research institutions and commercial companies involved in language and speech technology.

ELSNET aims to encourage and support fruitful collaboration between Europe's key players in research, development, integration, and deployment across the field of language and speech technology and neighbouring areas

ELSNET seeks to develop an environment which allows optimal exploitation of the available human and intellectual resources in order to advance the field. To this end, the Network has established an infrastructure for the sharing of knowledge, resources, problems, and solutions across the language and speech communities, and serving both academic

and industry. It has developed various structures (committees, special interest groups), events (summer schools, workshops), and services (website, e-mail lists, *_ELSNews_*, information dissemination, knowledge brokerage).

Electronic Mailing List

elsnet-list is ELSNET's electronic mailing list. E-mail sent to elsnet-list@let.uu.nl is received by all member site contact persons, as well as other interested parties. This mailing list may be used to announce activities, post job openings, or discuss issues which are relevant to ELSNET. To request additions/deletions/changes of address in the mailing list, please send mail to elsnet@let.uu.nl

Subscriptions

Subscriptions to *ELSNews* are currently free of charge. To subscribe, visit <http://www.elsnet.org> and follow the links to *ELSNews* and "subscription".

FOR INFORMATION

ELSNET

Utrecht Institute of Linguistics OTS, Utrecht University,
Trans 10, 3512 JK, Utrecht, The Netherlands

Tel: +31 30 253 6039

Fax: +31 30 253 6000

Email: elsnet@elsnet.org

Web: <http://www.elsnet.org>