elsnews 7.3

The Newsletter of the European Network in Language and Speech July 1998

Since its inception, ELSNET has regarded Linguistic Resources and Technology Evaluation as crucial foundations for the development of a thriving Language and Speech Technology. But in the early '90s it was also clear that, despite some important initiatives in Europe, the US ***Linguistic Data Consortium*** *and the* ***DARPA Human Language Technology Programme*** *had taken the pioneering roles in establishing a comprehensive framework for R&D in these areas. As* ***Antonio Zampolli*** *and* ***Joseph Mariani*** *note in this issue, however, there has been significant progress in building on the lessons of the US experience over the last few years, both in Europe and elsewhere; and the importance of these efforts was manifested in the huge success of the* ***First International Conference on Language Resources and Evaluation*** *that took place in Granada at the end of May. Not surprisingly, we felt it was appropriate to devote a whole (and extra-long) issue of Elsnews to the topics covered at LREC, and indeed to the conference itself.*

One of the key questions in Evaluation is: how far can the evaluation-driven methodology, which has proved so fruitful in the field of speech recognition, be generalized to other areas of language and speech technology? After a scene-setting review of the current situation in speech processing by ***Paul Taylor****, we examine how researchers are faring in such diverse fields as Machine Translation, Natural Language Generation, Grammar, Parsing, Spoken Dialogue Systems and Speech Synthesis, in an attempt to give a general picture of the burning issues.*

A second key question concerns the interplay between different interests in the efforts to build Language Resources. The relevant communities (the academic community, the industrial sector and government agencies) typically have different perspectives on issues such as collaboration (and competition), international access to national resources and corpora, and priorities with respect to targeting languages. These issues (and the corresponding role of funding) are all recurrent topics in this month's interviews and articles.

We hope you don't agree with everything in this issue. Please send us your comments, corrections and grievances: elsnews@let.ruu.nl. Have a good Summer.

Mimo Caenepeel, *ELSNews Editor*

*"I said I'd do anything to see the Alhambra" (***Melvyn Hunt,*** *Dragon Ltd)*

Note: In this issue the acronym **LR** is used to refer to Language Resources in general, including Speech Resources. Corpora, lexica, dictionaries, terminology banks and grammars are all examples of LR.
Other acronyms you will encounter a lot are **LREC** (The First International Conference on Language Resources and Evaluation), **LE** (Language Engineering), and **EC** (European Commission).

July 1998

elsnet

ISSN 1350-990X

500 people? You must be joking...

On the eve of the last day of LREC, we talked to ***Antonio Zampolli*** *about the background to the conference, the reasons for its enormous success, and the recent upsurge of interest in Language Resources.*

People working in the area of LR come from many different fields and often do not know each other. The original idea behind LREC was to bring them together, combining what they have to offer. I expected that an event like that would draw about 100 people; as it is, we have more than 500. That's amazing, if you consider that the conference focuses on LR and Evaluation, and that a general event like COLING, which covers all aspects of the field, usually has about 400-600 people.

There are a number of possible explanations for the unexpected success of LREC. I'm convinced that Granada has something to do with it. But Granada doesn't explain, for example, why the NSF [The US **National Science Foundation**] decided to pay for 50 people to come here.

Features that probably contributed to LREC's resounding success were the timeliness of the event in the general scientific and political context; the joint participation of Speech and NLP communities (a rather rare event); and the active participation of representatives of national and international authorities, funding agencies, major industries and SME;

How did we get to this point? Ten or fifteen years ago the prevalent approach in our field was mostly abstract. There was little interest in looking at real language: what people did was select an allegedly "interesting" example to study a particular form or property, or test some hypothesis with the computer. At one of the early conferences of the **European Chapter of the Association of Computational Linguistics** (ACL) — in 1982 I think — there was a presentation on the probabilistic tagging of a corpus. And some of the members of the ACL threatened to leave the Association because they felt that that kind of approach was not scientific at all, they were not interested in that kind of thing.

Today at ACL more than half of the papers are about empirical methods. So there has been some kind of change of paradigm over the past 10 years or so. A crucial point in this shift was a workshop in **Grosseto**, near Pisa, in 1987. There we brought together for the first time representatives from different fields — computational linguistics, AI, publishing, psychology, anthropology — who all recognised the need for developing lexica and corpora, the need for reusability, the need for standards and so on. In 1992 I introduced the term *Language Resources*, which had not been used till then, to underline the concept that resources are infrastructural. The term entered the literature, and ten years after Grosseto the time seemed ripe for another look at the state-of-the-art in the area, this time with a larger audience. But I hadn't expected 500 people...

The other big surprise at LREC has been the quality of the papers. There are still people who feel that LR is a field for 'workers in the street', so to speak, rather than for researchers and engineers. The quality of the work presented at LREC demonstrates how misguided such a view is. The problems that have been addressed here are now central to the whole field of Human Language Technology. After 40 years of computational linguistics we still don't have a parser that is able to analyse a real text; most parsers can analyse a few sentences, and then they stop. But the same research issues are at the centre of the interests of both LR and Computational Linguistics as a whole. And I believe that, since LR promote the data-driven approach, they can contribute to this effective capability of dealing with real language, in cooperation with the theoretical approach.

One of the things I'm slightly concerned about is the place that has been assigned to the development of LR in the **Fifth Framework Programme**. The EC has not explicitly reserved a cluster for LR, and that could be dangerous. On the one hand it may generate a duplication of efforts. And on the other, some important LR may in fact be non-sectoral, multi-functional and universal in scope; as such, LR is probably a field for cooperation between governments, industry and international agencies.

David Brooks, the Microsoft representative, gave an interesting presentation at LREC, and I agreed with what he said, except for his conclusion. At the end of his talk he classified languages in a number of categories, and said that Microsoft would develop resources for the first categories, and move on to the other ones progressively. But suppose you want to offer your knowledge and your commerce on the Web, in your own language — your language is not developed, because from the market point of view it is not important...

My own conclusion is: we need a policy to combine the market forces with some political principle. Microsoft is a commercial enterprise, of course; but the European Commission and the member states are not and must find a way to balance the market forces. I hope that 500 people speaking with the same voice are sending out a clear message in that respect..

FOR INFORMATION

Antonio Zampolli (pisa@ilc.pi.cnr.it, http://bibarea.area.pi.cnr.it/AREAEN/ilc.html) is Director of the **Institute of Computational Linguistics** in Pisa, and a member of the ELSNET Executive Board

ELSNews is published at the Centre for Cognitive Science, University of Edinburgh using Aldus Pagemaker™. It is printed on recycled paper by Lutton Press, Ltd.

Editors: Mimo Caenepeel, Mariken Broekhoven and Steven Krauwer.

Lay-out and production: Mimo Caenepeel

ISSN 1350-990X

FOR INFORMATION
Contributions to ELSNews, and address corrections, should be sent to:
elsnews@let.ruu.nl
Tel: +44 131 650 4594
Fax: +44 131 650 6626
Material for the next issue is due:
15 September 1998

Are we competing to cooperate, or cooperating to compete?

Mimo Caenepeel, University of Edinburgh

"Everybody's here," said one of the LREC participants at the lavish welcome reception on the first evening of the conference. And it seemed just about true, although a Spanish air strike prevented some of the speakers at the last moment from making it to Granada.

LREC was one of the largest planned efforts yet in the direction of integrating NL and Speech. It was an event on a big scale ("Like a fair," according to one participant), with a full, perhaps overfull, and varied programme that showed a good balance between NL and Speech, the general and the specific, and different formats. The papers covered a wide range of topics — systems, standards, LR, projects, theoretical issues and applications; they are bundled in two hefty volumes of proceedings (weight: 3.9 kg).

At the heart of conference a number of general issues, such as: Who takes care of multilingual resources? What are the priorities and most urgent needs? What is our vision for the future? To begin with, there seemed to be general agreement on a number of things: the need to protect and provide LR for the lesser-used languages, for instance, to make sure they do not fall behind in becoming part of the electronic age; the undisputed importance of standards; the fact that it makes sense to collaborate and share expertise, and to create common corpora.

But in the course of the three days, differences began to emerge as well, particularly in the viewpoints of the different communities represented at the conference (the academic community, the commercial sector and government entities) with respect to funding priorities. Of course LR should be developed and validated, but It Costs. And the industrial perspective, in the words of **F. Kunzman** (**IBM Europe**, Germany), "comes down to the money issue: you have to make money for your corporation." This affects policy with respect to sharing LR: "We will buy ELRA resources, but we are not willing to share ours", said **Nils Lenke** of **Philips**, Germany. **David Brooks** of **Microsoft** put this further in perspective when he outlined the Microsoft approach to funding. Microsoft prioritises languages depending on the numbers of computers they sell in the relevant countries. Brooks' presentation listed the different categories of languages they target and fund, with English in the top category, 'major' CEE languages in the second one, and so on. Several of the contributors to this issue of *ELSNews* comment on Microsoft's approach.

Obviously market forces and industrial needs cannot be ignored. But we should not forget the motivations of the research community, the needs of the language community at large, and the principle of democratic access to a coordinated infrastructure. Many people at the conference felt that governments, and the European Commission in particular, need to play a balancing role in this respect. The official response to this was less definite: **Roberto Cencioni** declared that he was prepared to put 30M ECU on the table "if the heavyweight players are prepared to do the same." (**Antonio Zampolli** comments on the EC's approach to funding LR on the previous page).

In the closing session of the main conference, there were summaries of the general outcome of papers in different areas Not surprisingly, work on Resources in the Speech Area came out as most explicitly successful: **Harald Höge** emphasized the rapid progress in this area, but also talked about the urgent need for further development and standardization of SRL. In the area of Written Language Resources there were, according to **Nicoletta Calzolari**, no real new trends, but the number and quantity of submitted papers was a surprise, and an indication of the global level of maturity of the field. **Joseph Mariani** spoke of projects in the area of Spoken Language Evaluation (cf. p 13 of this issue), and discussed some recent results on discrepancies between technology-based and user-based evaluation projects in the area of Spoken Language Evaluation (see also **Gerrit Bloothooft**'s piece on p 8). **Bente Maegaard** concluded that in the area of Written Language, too, Evaluation as a science is becoming mature and standards are emerging, although a lot more work is needed. **Khalid Choukri**, finally, reflected on the industrial involvement in LREC which was not, he emphasized, just an academic conference.

Stealing the show: some of the major LREC players on the last evening of the conference.

There were many lighter moments at LREC, and at the grand banquet on the last evening some participants revealed a (suspected or unsuspected) talent for dramatization and spectacle. In general, it seems clear that we could all benefit from more collaboration and more flamenco dancing. Or, in **Maghi King**'s words: If everybody is talking the same kind of language, it is much easier to talk to one another.

Comment

Point of View: Marc Blasband

I have returned from LREC with a message that is simple but strong:

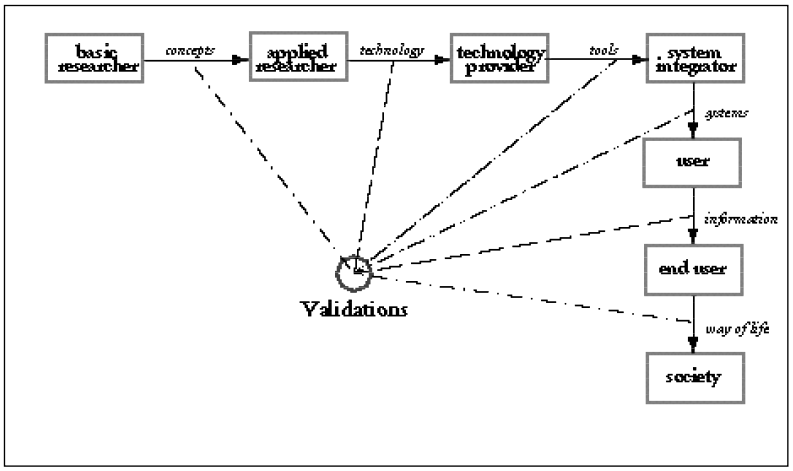
The whole LE process must be validated from basic research to social impact

So far the LE community has paid attention mostly to the horizontal dimension of the figure on the right: the validation of the technology. Now the vertical dimension needs to be thought about and researched, to rectify the balance and ensure a comprehensive validation of the entire field.

The advantage of such a comprehensive view of validation is obvious: it would allow us to measure the impact of our work on society as a whole, and justify the years of work that are now culminating in positive results. But it also means that additional work is needed. We will need to define the relations in the figure above, guess what criteria the components must satisfy, and determine how to measure and compare them. We should also indulge in some science fiction, and think about the changes in society the technology might bring.

This comprehensive concept of validation can also be used from the opposite side: we can reverse the arrows on the figure to determine what performance is required at a given level to allow a particular result at the next step. Here we have a typical waterfall model, and as for most waterfall models we need salmons to swim upstream (in this case to percolate the measures from the latter steps to the previous ones).

On the basis of a way of life we could sequentially specify the properties of systems for end-users, tools, technology and concepts that will ensure the performance required at every link in the chain. For translation, for example, we could specify the requirements for different areas: translation for people who travel, people who use machines, or people in conversation. For each of those we could specify a domain size, an acceptable error level, and a required processing speed. From these attributes and values we could determine requirements of the system, tools, technology and concepts, and plot their development realistically for a few years.



The validation process would involve reviewing these relationships on a very regular basis, taking all the new developments into account. Clearly no one will have the foresight to predict all possibilities of the technology and the ways to achieve the results hoped for, or the foresight to include them in a research programme. So a lot of care will need to be exercised here.

It is obvious that such a plan would not resist the push of reality, and would not be valid for more than a year or two. However, I feel that it could provide focus and direction to the endeavours of the whole community. To achieve that, it must be small and very flexible, and seen by everybody as a tool to position the goals and the results of every step.

Up to now, the field has progressed in a bottom-up way. The top-down approach I'm suggesting here will only be successful if it is very flexible. History has shown that a planned approach to innovation is rarely successful. But measuring progress against visions of a possible future could focus research and indicate what we might achieve. It should also show under which conditions we could reach these results.

When approached like this, the validation process has a double function: to determine the performance of one aspect of the technology, and to check whether this will be sufficient for a particular usage downstream. In other words, it would add some demand pull to the technology push that has brought us the success we have now.

Marc Blasband shares a joke with Steven Krauwer at LREC.

Evaluation

On the next seven pages we give an overview of current Evaluation practices in different areas of NL and Speech. What are the main challenges in developing a successful Evaluation methodology for each area, and what are the most promising avenues for future work? ***Paul Taylor*** *kicks off with a summary of the state-of-the-art in Speech Recognition.*

Speech Recognition

Paul Taylor*, CSTR, University of Edinburgh*

Objective evaluation criteria have been used in Automatic Speech Recognition (ASR) research since serious work started in this field. One could say that the use of such evaluation criteria is a defining character of ASR, in that there exists a simple and uniformly-agreed metric for evaluation of speech recogniser output.

In essence, ASR systems are assessed on open testing data, that is, data which the system has not had access to during training. The words in the test data have been manually transcribed, and the system is evaluated by measuring how closely the automatic and manual transcriptions match.

The fact that the test data has not been seen by the system is crucial. Given enough free parameters, it is easy enough for any system to memorise all the examples in the training data, and hence achieve a perfect score. Such an approach is not possible with independent test data, and hence systems can only perform well if they have captured the important generalisations in the task. However, this in itself is not enough — during the development of a system for a particular domain, researchers might run thousands of tests on the test data, and it is inevitable that the idiosyncracies of the test data will be picked up too. To combat this, two test sets are often used: a development set for day-to-day use, and an evaluation set that is only used at the end of a substantial period of research.

Speech recognition funding, particularly in the US, is often structured around solving problems in particular domains, with all research groups using the same training and test data. Over the years various tasks have been proposed, and groups are funded to do research in these domains. Periodically (often every 6 or 12 months), new unseen test data is released, and research groups run their current recogniser on this and submit the results. A table of results for all the groups taking part is then published.

This type of evaluation has its fans and critics. The fans (often funding agencies) point to the remarkable improvement in ASR performance over the last twenty years, from systems capable of recognising a few hundred words of fluent speech from a single speaker to systems today which can recognise tens of thousands of words of spontaneous speech from any speaker. The scientific breakthroughs responsible for this are numerous, but many think the main reason for the improvement lies in easy evaluation. Within a research group it is possible to test out new ideas quickly and testing between groups makes the field fiercely competitive.

Two main criticisms are often encountered. The first is that the nature of the test is unrealistic: it is absurd to treat all words, and hence all word errors, equally. For instance, in the sentence "I am definitely not guilty", a deletion of the word "not" may have much more serious consequences than the misrecognition of other words. Some have proposed that recognisers should be assessed in terms of their ultimate purpose, as the ability of the whole system to perform the given task is all that matters in the end. For instance, when AT&T developed a system for handling telephone enquiries in a department store, the system was assessed only on whether it connected the caller to the right person, not on any measure of word accuracy. While this simulates what one would want out of a system in terms of performance, it is more complicated than comparing transcriptions, and less adaptable to other domains. Proponents of transcription evaluation have countered saying that although ulimate whole-system performance is important, improvements in word accuracy will lead to this anyway, and so it is fine to deal with that alone during development.

The other main type of criticism claims that the evaluation of speech recognisers has become obsessive, and has led to research groups pursuing only short-term solutions. When competitive evaluations are performed every six months there is little time to develop radical new ideas. Such critics also point to the fact that the vast majority of ASR systems today employ more or less the same **hidden Markov model** (HMM) technology. There is no consensus on whether ASR research is running into a dead end; and while it seems unlikely that HMMs as currently used will provide the final solution, it is still the case that year after year the top-performing systems get noticeably better. However, there does seem to be a feeling by most an obsessive attention to system coparison is counterproductive, and that researchers should be given more freedom in pursuit of new techniques.

A reasonable question to ask is whether the success of testing in speech recognition can be adapted to other areas in speech and language processing. Unfortunately it seems to be an idiosyncratic property of the ASR task that such an easy evaluation as word-transcription comparison is available — in other areas no one simple score seems to capture the performance of a system adequately. For instance, how can one compare parsers which don't use the same syntactic categories? Hopefully, fair metrics will be devised for other problems, and then the basic methodology of ASR testing — open test data and competative evaulation — can be adopted elsewhere.

FOR INFORMATION

Paul Taylor (Paul.Taylor@ed.ac.uk, http://www.cstr.ed.ac.uk/~pault/) works at the **Centre for Speech Technology Research** (**CSTR**) at the University of Edinburgh. His main research interests are in the areas of speech synthesis, speech recognition, prosody and phonology.

Evaluation

Machine Translation

Eduard Hovy, *Information Sciences Institute of the University of Southern California*

Like a good mystery novel, MT evaluation is an old topic that refuses to be forgotten. Even when you know the answer, you still feel like getting into it again. And so many people do, according to **Yorick Wilks**, that more has been written about this topic than about MT itself! Yet after 50 years there is still no standard test, no internationally accepted rating system, and little agreement about what exactly should be measured.

In an attempt to develop a definitive method, **DARPA** in 1990-94 hosted a series of four MT Evaluation competitions, pitting human translators, research systems of various kinds, and commercial systems against each other. In these evaluations, each translation was judged in three ways: with respect to its grammatical fluency, its adequacy (how much material was left out), and its comprehensibility (what content the reader could glean regardless of other inadequacies). This exercise was phenomenally expensive (the last one in the series, involving 18 entrants, took over three months and cost over $400,000) and its results were only mildly informative. By trying to be all things to all people, the results were neither focused enough on system details (system-internal or glass-box) to help system builders pinpoint errors, nor focused enough on users' needs (functional or black-box) to allow potential users to decide whether they would be interested in buying the system for some application.

If there is one thing this researcher has learned, it is that no single measure will do. There is no one villain. As in a good Agatha Christie mystery there are lots of potential murderers, and usually more than one of them conspire to create the mayhem. Multidimensional evaluation techniques are, in my opinion, a necessity. It is imperative to enable the user to assemble his or her own evaluation, by selecting from a smörgåsbord of measures and then combining the scores into a few simple overall numbers. Inevitably, the user's selection will depend on his or her intended use for MT.

This implies that MT evaluation researchers should concentrate on creating, coordinating, and cross-calibrating whole sets of tests, on a variety of dimensions. A few such evaluation schemes have indeed been proposed and tested (Nomura 92 ; Mason & Rinsche 95).

These kinds of multidimensional schemes are a step in the right direction. But they do not yet fully meet the user's needs; the user cannot vary the relative importance of any single aspect with respect to the others. The kind of MT evaluation I have in mind is a multidimensional one, where the various dimensions are organized into a taxonomy of ever-increasing specificity. An appropriate evaluation measure, of appropriate delicacy, is associated with each level of each branch of the taxonomy. The user is then free to select the desired level of delicacy along each branch of the taxonomy, apply the evaluation measure found there, and (if desired) propagate the resulting scores back toward the root of the taxonomy. An example of such a taxonomy of evaluation measures can be found in the paper I presented at LREC.

Obviously, multidimensional evaluation is not particular to MT. It can (and should) be applied to all complex Natural Language Processing endeavours, including Information Retrieval and Text Summarization. Then we will be much better able to appreciate the mysteries inherent in that wonderfully complex thing, language.

References

Nomura, H. *JEIDA: Methodology and Criteria on Machine Translation Evaluation* (JEIDA Report). Japan Electronic Industry Development Association, 1992.

Mason, J. and A. Rinsche: *Translation Technology Products.* OVUM Ltd., London,1995.

Hovy, E.: *Creating useful Evaluation Metrics for Machine Translation*. Paper presented at LREC, Granada, 1998.

FOR INFORMATION
Eduard Hovy (hovy@isi.edu, http://www.isi.edu/natural-language/nlp-at-isi.html) is Project Leader of the Natural Language Group at the Information Sciences Institute (ISI), University of California. His research interests include automated text summarization, machine translation, text planning and generation, semi-automated construction of large semantic resources, and multilingual information retrieval.

*The **Palacio de Exposiciones y Congresos** in Granada, venue of the LREC conference*

Natural Language Generation

Evaluation

Robert Dale, *Microsoft Research Institute, Sydney*

"Natural language generation" covers a great many things, just as the terms "natural language processing" and "natural language understanding" do. Nobody would say: "I evaluate NLP systems"; similarly it doesn't make sense to say "I evaluate NLG systems". The problem is too big for that and, just as in the case of NL understanding, we need to determine how to break it down. One of the things that makes this especially hard in NLG is that, as I think **Yorick Wilks** once put it, the problem of natural language analysis is somewhat like counting from one to infinity, whereas in language generation you're counting from infinity to one. So you know roughly where you want to get to — a text — but you don't really know where to start from. One response to that is to say: "Let's take input from sources (like databases or expert systems) that already exist." But those knowledge sources rarely provide the kinds of distinctions that you need to motivate the kinds of variations researchers want to explore in NLG systems. The other problem in generation is: how do you assess what you end up with? It's hard enough to evaluate text written by humans. For NLG, you start out with questions like: Did the system generate as coherent a text to communicate the information as it might have done? Did the system put the right amount of information into the text, no more and no less? Those are incredibly hard things to assess.

Some people have tried to use comprehension tests: you have the machine generate some text, you get a human subject to read the text and answer questions about it, and you try and assess the quality of the text on the basis of that. But there are so many variables in such an experiment that it's hard to know what you are really measuring; plus it doesn't give you a direct insight into what you would have to do to the system to improve it. If you try that kind of comprehension test on a human-authored text, you can at least rely on the author's self-knowledge, intuition and capabilities to work out what she would have to do to get it right the next time.

Many aspects of NL understanding are difficult to evaluate too, of course. But evaluation tasks in that area all seem to have a particular quality, which is to do with canonicalisation, normalisation and the restriction of results: there is a clearly defined target. For example, you've got a noun phrase, you want to know what it co-refers with, and you know there's only a finite set of possibilities. It's hard to see what the correlates of those kinds of tasks would be in generation, even for such narrow evaluation scenarios: there is, by and large, no one right answer when it comes to generating a text. There are many texts that serve the same purpose, we know that some are better than others, but how you actually quantify that is just not clear.

We may be able to draw inspiration from MT or text summarisation, two other domains where textual output is the issue, to determine what to evaluate. But there is a fundamental difference between MT and text summarisation on the one hand and NLG on the other, which comes back to the one-to-infinity/infinity to-one problem. In the case of MT and text summarisation you've got input text: in both cases you could replicate the relevant tasks using a human subject, and compare the results. But you can't really give a human subject the input to a generation system and say "Go generate some text from this knowledge base and see if you can do it better than a machine can".

There may be some specific things where you can start to make more headway. Take the generation of pronouns as a form of reference: you could say that a system that is good at that task will generate pronouns when they're not ambiguous, for instance. One problem here is of course that normal reference resolution is not done in a vacuum — people's world knowledge makes a difference to the ease of interpretation. So it's not as if there's an objective standard. But if we ignore that problem for the moment, and take on board the suggestion that one can reasonably talk about pronouns either being appropriate or not appropriate in a context, we could manufacture an experiment for that, and indeed this has been tried. So there may be some tasks like that where you can start to move in the direction of a quantifiable metric.

Some exploratory work that **Chris Mellish** and I reported on in Granada starts from work on defining architectures for generation systems that I've been doing with **Ehud Reiter**. The way we have been looking at it so far is to come to the problem not from a black-box evaluation perspective but from the point of view of system architecture. An NLG system will have some component that works out what the content of the text will be; there might be some other component that works out the structure of the text, and a component which decides how to lexicalise concepts, and so on. So we might ask: what do those components add to the output text, and can we begin to evaluate on the basis of that? That is one way of getting a better grasp on the problem. It's still not trivial, because different people have different views as to what the subprocesses involved are. It's like a move towards a glass box evaluation, except there is no agreement on what the modules are of what one is evaluating. But we figure that by decomposing the task in this way and looking at the contributions of individual processes in generation, we may start to make some headway. But it's early days, still. From where I sit at the moment, my intuition is that there's something fundamentally different about evaluating NLG systems, as compared to the rather specific evaluation scenarios we see in NL analysis work.

FOR INFORMATION

Robert Dale (Robert.Dale@mq.edu.au, http://www.mri.mq.edu.au/~rdale/activities/) is the Director of the Microsoft Research Institute (MRI) at Macquarie University, Sydney, Australia. He heads the Institute's Language Technology Group. Over the years, his research interests have been focused in three major areas: the generation of referring expressions; the use of shallow approaches to intelligent text processing; and the role of visual elements of language delivery in communicating meaning. In his more recent work, these strands have been brought together in exploring how natural language processing techniques can be used on the WWW.

Evaluation

Parsing

John Carroll, School of Cognitive and Computing Sciences, University of Sussex

Parsing is an essential part of many larger applications, such as information extraction systems, which have gained in importance over the last few years. In such applications the parser and grammar are often central components, and achieving good results relies on being able to select an appropriate parsing technology, and determining and improving weaknesses in an existing parser/grammar. Reliable parser/grammar evaluation methods are therefore vital.

There are two different objectives for evaluation in parsing (**Srinivas** 98). *Intrinsic evaluation* refers to the evaluation of particular systems in order to monitor their development and diagnose areas of weakness. A repertoire of techniques exists for this type of evaluation, including measures of coverage and correctness with respect to parser/grammar-specific reference corpora. Task-based measures can be also used, for example to determine whether the performance of the larger application improves when the parser/grammar is changed in a certain way.

Extrinsic evaluation is concerned with establishing an evaluation method for comparing the accuracy of different parsing systems with respect to an (annotated) reference corpus. The extrinsic parser evaluation method which is currently most widely used is the **Parseval** scheme. Parseval requires the reference corpus to contain a bracketing for each sentence. Parser output is scored on the basis of the number of bracketings that match the reference (giving bracket recall and precision figures), and also the number of crossings, indicating the degree to which the two sets of bracketings are mutually inconsistent.

But the Parseval scheme has a number of limitations and drawbacks, including a commitment to a particular style of grammatical analysis, an oversensitivity to certain innocuous types of misanalysis, and an occasional failure to penalise common types of more serious mistakes. Alternatives discussed at the recent **Workshop on the Evaluation of Parsing Systems** include revised measures within a modified Parseval scheme, and measures with respect to a dependency-style annotation of the reference corpus. These proposals were debated with some vigour at the plenary session of the workshop, and there will clearly be follow-up work building on the workshop results.

Reference

Srinivas Bangalore, Anoop Sarkar, Christine Doran & Beth Ann Hockey: *Grammar & Parser Evaluation in the XTAG Project.* In: Proceedings of the Workshop on the Evaluation of Parsing Systems, LREC, Granada, 1998.

FOR INFORMATION

John Carroll
(johnca@cogs.susx.ac.uk, http://www.cogs.susx.ac.uk/lab/nlp/carroll/carroll.html) is an EPSRC Advanced Research Fellow at the School of Cognitive and Computing Sciences, University of Sussex. His research interests are in the areas of large-scale grammar and lexicon development, practical NL parsing, and robust analysis of unrestricted English text. The projects he is involved in include **SPARKLE** (Shallow PARsing and Knowledge extraction for Language Engineering); Analysis of Naturally-occurring English Text with Stochastic Lexicalized Grammars; and **PSET** (Practical Simplification of English Text)

The **Workshop on the Evaluation of Parsing Systems** was organised in collaboration with researchers from the EC Language Engineering projects SPARKLE and ECRAN. The ten refereed papers presented are published in a set of proceedings, which is available as a University of Sussex technical report (http://www.cogs.susx.ac.uk/cgi-bin/htmlcogsreps?csrp489).

Announcement

EAGLES Handbook of Current Evaluation Practices

The **EAGLES Evaluation Group** is putting together a **Handbook on Evaluation methodologies** with the title above. The body of the report will concentrate on the EAGLES attempts to build on the ISO 9126 standard (evaluation of software) in order to produce a framework for designing evaluations of LE systems. There will also be substantial appendices reflecting current evaluation practices in LE. We are anxious that this part of the report should cover as wide a ground as possible, and that the report as a whole should reflect the current state of the art in evaluation throughout the LE community. We would therefore be very glad to hear from anyone who is prepared to make a contribution to the report, for instance through:

- an offer to read and comment on draft versions of the report — existing papers or reports on evaluation within language engineering projects
- existing papers or reports on evaluation of language engineering products or systems
- commentaries on previous EAGLES evaluation work
- commentaries on evaluation techniques and methods

Or, to put it more succinctly, if you think you may have something to offer us, please get in touch! You can contact us through our web site at http://www.cst.ku.dk/projects/eagles2.html, where you can also find links to previous EAGLES reports as well as other material on evaluation and on EAGLES activities.

Grammar

Klaus Netter, DFKI, Saarbrücken

Given the importance that is typically attributed to grammar development in computational linguistics and language engineering, it is really surprising how comparatively little is happening in that area in terms of systematic and comparative evaluation. This may have to do with the fact that there are relatively few sizeable grammar fragments (outside the English language), but it is obviously also due to a lack of suitable reference data as the basis for evaluation (again leaving aside English data, as for example in the **Penn Tree Bank**).

To evaluate a grammar component by itself (and not embedded in an application) is clearly not an easy task, as it will practically always go together with an evaluation of a parser (which is probably why **John Carroll**, in his contribution on parser evaluation, prefers to talk of parser/grammar evaluation). Among the criteria to take into account are parameters such as coverage and speed, as well as the number of readings assigned to a string, all of which will also depend on the processing components. While coverage is more obviously a criterion for measuring a grammar, it is certainly not a straightforward criterion, since it also has to be seen in combination with the structure assigned as a parse result. Comparing such structures across different grammar frameworks so far has not been achieved in a fully satisfactory way. Speed can be a relevant criterion, if all other parameters are kept unchanged, so one can measure what methods in specifying a grammar are more effective. Again, this is not uncontroversial, if one keeps in mind that the effectiveness of a particular style of grammar writing may depend on, and also go together with, a specific method of processing or compilation.

The most difficult evaluation criterion is probably the number of readings assigned to a string, since here semantics could play a crucial rule. Do you get false readings, spurious readings and how do you define such readings? On the one hand, syntactically well-formed structures which would have to be translated into rather marginal or not quite so salient semantic interpretations count as failures, and where exactly should the borderlines be drawn? On the other hand, should underspecified structures compatible with different readings count as misses? How does your grammar perform on negative or ill-formed examples? Does it assign a reading to a negative example that was not intended? Again the interference with the parser plays a role, since a robust parser shouldn't break down on bad examples, but it shouldn't pretend to have found a grammatical reading either. Ideally it should be able to give you a fragmentary result: tell you that although parts are well-formed, altogether the reading doesn't make sense, and it should be able to say why.

Some of these problems can be solved to a limited degree if the parameters can be investigated very systematically. Some of the current work on Grammar Evaluation is therefore based on *test suites*, i.e. corpora of systematically constructed grammatical and ungrammatical test items which can act as a standard or benchmark for measuring such performance criteria. Such test suites allow for very precise testing and diagnosis of grammars, by providing a controlled environment; in particular, they can help to explore the coverage by testing the ability of a grammar to distinguish well-formed from ill-formed items. *Diagnostic evaluation*, which helps to identify deficiency of a grammar, and *progress evaluation*, which can record the performance changes across different versions of a grammar, are probably best carried out on the basis of such test suites. We had some experience of test suite based grammar evaluation in combination with the TSNLP and DiET projects, which looked quite promising. Still, I am afraid that we are still some way away from broadly accepted evaluation and measurement criteria and methodologies as well as from large-scale comparative grammar evaluation.

Of course, over recent years the general interest in rule-based processing on the basis of full-fledged grammatical analysis has diminished, with statistical methods yielding quite acceptable results for certain applications. However, it appears that the pendulum is about to swing into the other direction again, with a revival of the rule-based approaches, in particular where they are combined with statistical methods. In the course of this, grammar evaluation is also becoming more interesting again. The work of the **XTAG** group at UPenn, combining a neatly specified **TAG** grammar with methods of supertagging and evaluating this agains the **Wall Street Journal**, is just one example. There is some substantial work on **HPSG** carried out in the **Verbmobil** project, where English and German grammars are evaluated against the **TSNLP** test suites and the Verbmobil corpus. Even applications which traditionally involved little if any grammatical knowledge, such as information retrieval, are opening up more to the use of rule-based methods, be it only for shallow processing. For example, in the **Twenty-one** project, which offers an online multi-lingual retrieval engine, the evaluation of the quality of the phrasal chunking was taken quite seriously, as it is intended to support the user in judging the relevance of a document. (See also the contribution by **John Carroll**, on the opposite page, for further examples.)

In short, I expect that with the increasing number of applications involving not only shallow but also deeper grammatical processing, the need for evaluation will be there again and it will increase, in very much the same way as evaluation in the speech community has already become an issue with some quite serious commercial implications.

FOR INFORMATION

Klaus Netter (netter@dfki.de, http://www.dfki.de/~netter/) is the deputy head of the Language Technology Lab at DFKI, where he has worked as a researcher and project manager since 1990. His current focus is on application-oriented research in the areas of multilingual multimedia information retrieval, as well as the testing and evaluation of NLP components and applications.

For more information on the projects mentioned, see
TSNLP: http://lt-www.dfki.de/tsnlp
DiET: http://www.dfki.de/pas/f2w.cgi?ltp/diet-e
XTAG: http://www.cis.upenn.edu/~xtag/
Verbmobil: http://www.dfki.de/verbmobil
Twenty-one: http://www.dfki.de/pas/f2w.cgi?ltp/twenty-one-e

There will be a project update on **DiET** in the next issue of ELSNews.

Evaluation

Spoken Dialogue Systems

Gerrit Bloothooft, Utrecht University

Every day, 1500-7000 people in the Netherlands consult the automatic railway information service. This figure shows that spoken dialogue systems have passed the factory acceptance test. But we also know that the technology behind them is far from perfect. And how well are the systems received by their actual users? Which parts need to be improved, and in what way?

Evaluation might provide answers to these questions, and they were the focus of many of the presentations at LREC. A general distinction emerged between two mainstream types of evaluation: a caller-oriented and a technology-oriented one. It was concluded that both types are essential, since it is often not clear how to interpret the results of one type of evaluation in the framework of the other. On the other hand, none of the presentations actually described a combined test of both subjective and objective measures, and more research on this topic seems warranted. Closest came a pilot analysis of the Dutch railway information system, where the technological evaluation claimed a 94% success rate, while only 66% of the users thought they had completed the task (and only 30% without error repair). This apparent disparity could be explained completely, however, by the way the numbers were interpreted: 14% of the users had accepted information which was wrong (asking for a train to Rotterdam and getting the answer for Amsterdam, for instance), but this was considered a successful task completion in the factory acceptance test; 15% of the users were just playing around with the system (for example, asking for station names with minimal pairs such as 'Maarn' and 'Baarn') and were excluded in the original count.

The results of the ELSNET Olympics showed that the three main dimensions of users' judgments are general appreciation (including task completion and error recovery), functional capabilities, and the system speech. The number of turns, the time span of the dialogue and the increase in the use of numbers were mentioned by **Philips**' **Christian Dugast** as objective measures of success. A comparison of the Swiss railway information system (+41.1.570222) from **Philips** and a French **DTMF** (Dual Tone Multi Frequency) system showed that the spoken dialogue system was three times faster (40 seconds versus 2 minutes), which is a very strong commercial argument in favour of spoken dialogue systems.

Alongside user evaluation, component assessment can be used successfully to identify the technological needs for improvement. For the MIT **Jupiter** weather information system, for instance, evaluation metrics were applied to each component, and a decrease of word error rate from 35% to 8% during the first year of operation (6500 callers) and a current parse coverage of about 99% provided a clear measure of success. The general expectation at LREC was that the dialogue management component would be hardest to assess. But in the Dutch case users tend not to converse with the system to begin with: when a system error occurs they do not exploit (or do not suspect) possibilities for error recovery, but hang up instead or ask for the operator. Both man and machine still have a lot to learn.

FOR INFORMATION
Gerrit Bloothooft
(Gerrit.Bloothooft@let.uu.nl) is researcher and lecturer at the Utrecht Institute of Linguistics OTS. His research interests include speech recognition, voice quality measurements and singing.

Announcement

Euromap network promotes language technology

The **Euromap** network will launch its language technology awareness programme in the summer of 1998. Euromap is an EU-wide network set up to provide information about the language technology field in general, as well as to disseminate information about the forthcoming Fifth Framework Programme, and opportunities for language technology research and development funding. Euromap nodes (called *National Focal Points*) have been established in all Member States.

Each network node can:

- provide information about language technology suppliers, researchers and developers in the member state, and through its links with partners, elsewhere in Europe
- provide publications and information about language technology projects supported by the European Commission
- give presentations and workshops on language technology, and its application in the information society and the digital economy
- provide information and tutorials to potential partners in EU-funded language technology projects
- provide a partner-search service for researchers, developers, suppliers and/or users who may wish to participate in projects.

For more information, contact the project co-ordinator:
Bente Maegaard
bente@cst.ku.dk
Tel +45 35 32 90 74
Email: bente@cst.ku.dk

Center for Sprogteknologi
Njalsgade 80
DK-2300 Kobenhavn S
Tel: +45 35 32 90 90
Fax: + 45 35 32 90 89

Speech Synthesis

Evaluation

Louis Pols, University of Amsterdam

Although Text-to-Speech (TTS) synthesis development and evaluation is only a relatively small aspect of Speech Technology, it is a vital part of any spoken dialogue system. It will also continue to be a test case for our (lack of) knowledge about all aspects of speaking, from text interpretation to voice realization.

The papers on Speech Synthesis presented at LREC fell into two groups: some were concerned with improving performance in a specific language, others focused on aspects of system evaluation. Among the languages that got attention were French, Slovenian, Dutch, Japanese and English. There was a diversity of approaches, both database-oriented and rule-oriented ones. Specific aspects of system evaluation include grapheme-to-phoneme conversion, prosody, and speech quality. When it came to evaluating the performance of complete dialogue systems, attention for the speech output part was generally small.

One interesting new avenue of investigation in this area concerns the use of the Internet for direct TTS access with any text imaginable. Another interesting new direction is the use of large text corpora (such as newspaper texts, telephone directory entries, raw e-mail messages, weather reports, and so on) as an independent source for text input for synthesis evaluation. At the forthcoming ESCA Synthesis Workshop, many different systems in several different languages will be tested according to these principles.

FOR INFORMATION

Louis Pols (pols@fon.let.uva.nl, http://fonsg3.let.uva.nl) is professor in Phonetic Sciences at the University of Amsterdam and chairman of the TTS evaluation committee that coordinates system evaluation at the forthcoming Australian Synthesis workshop. His main research interests are in speech perception and in evaluating the performance of speech technology systems and components.

To find out more about Internet use for direct TTS access, visit http://www.ldc.upenn.edu/ltts/.

The **ESCA Synthesis workshop** will be held at Jenolan Caves, Australia, November 27-29, 1998 (preceding ICSLP'98). For more details, see http://www.itl.atr.co.jp/cocosda/synthesis/3rd_ws.html

Announcement

Conference on Advanced Computing in the Humanities

The **SOCRATES** thematic network project on **Advanced Computing in the Humanities** (**ACO*HUM**) will be organizing a conference in Bergen (Norway) on September 25-28 1998. The conference will provide a forum for discussing the role of computing in a wide range of disciplines, ranging from natural language and speech to historical databases and digitized art. The preliminary programme has the following sessions:

- Reshaping humanities education in a digital age — sharing content across institutions
- Constructing digital sites
- Tools for the humanities
- Course development on the Internet
- Using digital text resources
- Cross-border curricula
- Transnational networking
- Scenario's for the digital classroom
- Curriculum innovation: impact on disciplines

There will be research demos and an industrial exhibition; proposals for those can still be submitted. The last day of the conference will be reserved for workshops, and a special workshop is planned on coordination of activities in NLP and Speech.

The conference is aimed at academic staff, planners and innovators, project leaders a. people working in libraries, museums and so on, publishers and other content providers. The programme committee is chaired by **Koenraad de Smedt** and **Daniel Apollon**, both at the University of Bergen.

For more detailed information, please consult the conference web site: http://www.futurehum.uib.no; or contact the conference secretariat by email (futurehum@uib.no) or phone (+47 5558 8008).

A strange friendship

There were a striking number of US participants at LREC, many of them funded by the NSF (cf. Antonio Zampolli's remarks on p 2). ***Mimo Caenepeel*** *spoke to* ***Judith Klavans,*** *Director of the* ***Center for Research on Information Access*** *at Columbia University, about collaboration, competition, and the role of industry and academia.*

ELSNews: *What are in your view the main differences between the US and Europe in terms of LR policy, both at government level and within the relevant industrial and academic communities? Is there scope for trading expertise or LR in particular areas?*

Klavans: Europe and the US are in strategically different positions. The US has only one (or two) national languages, and it would not make sense for us to collect only English data, because of obvious strategic disadvantages. The US community has done a lot more work than the EU on Asian languages, since we have large numbers of Asian speakers, and we have very close links. Covering corpora on a particular national language doesn't have to be done within the country where the language is spoken, as long as you have experts who know the language. The EU is in a different situation, because it has 15 official languages, and many other unofficial ones. So what we come with from the States is more experience in dealing with Asian languages, and what you come with is a more pressing need, and more experience in dealing with more languages. We both have something to offer to each other.

There is a sense of competition, of course. The **LDC** (the US **Linguistic Data Consortium**) started in 1988; the **European Data Collection** effort began in 1990. At that point in time the computational linguistics community in Europe was beginning to grow larger, stronger and better-funded than the US one, so there was an increasing sense of competition. It also seemed to be the case that LR were a core part of the US effort, while the EU really had to struggle to get funding in that area. This shifted when the EC began to realise that LR were going to be a strategic issue, and started pouring money into it. The result of that is that the field has grown more quickly in Europe recently.

It's a strange friendship, and I think we have some way to go before we overcome that. I suspect the policy area is where it needs to be done.

I would also say that both in the US and in the EU the research community and the government could be paying more attention to what is going on in industry. There is not enough awareness within these communities of the fact that industry is pushing a lot of multilingual corpora that never get outside the company. When I was working for **IBM**, for example, the company was targeting 10 or 11 national languages, had a huge terminology bank, and was doing a lot of work on the (terminological) translation of manuals, with whole databases being developed of things like the most frequent translation for a particular kind of term in a particular kind of document. None of that work has ever been published outside the business community.

In his talk at LREC, **David Brooks** of **Microsoft** was saying that Microsoft prioritises those languages which are spoken in countries where they sell the most computers. I think that talk was a red flag for us, the academic community, because the corpora we build are small compared to what already exists in manuals. Businesses will do French before they do Thai. So it is incumbent upon us, as keepers of language, to make sure that we build LR where industry won't do it. I think that's a funding priority for governments.

ELSNews: *It was emphasized time and time again at LREC that collaboration — between people, communities and countries — in the area of LR is crucial. What could be done to encourage this? What are the main barriers?*

Klavans: Let me give you the experience from the perspective of the **Multilingual Evaluation Access** working group which I'm involved in. There are two working groups, one from Europe and one from the US; both have six people. The funding is only for travel. The exact same proposal with exactly the same goals was submitted to the parallel directorates; and as a result there was not that sense of pulling in two different directions, no feeling of "you may be taking away from my funding"; none of that happened, it was truly coordinated.

That kind of policy, I think, fosters real collaboration. I don't want to be naive about power versus collaboration, but I do believe that with the proper financial encouragement it can work. Those who do like to collaborate really respond to that kind of structure, and if people are not naturally collaborative they adjust. Our respective government agencies could do a lot to make that happen. I've seen it in these working groups, it is really quite remarkable.

Money is a great enabler if it is structured correctly, and if it is shared among groups which are required to cooperate in order to get funding. Another example: on the **Digital Library** program in the States, the PIs (Principal Investigators) meet once every 6 months, and if you want to get supplements to your proposals you have to do something collaborative. You can't do it on your own: no collaboration, no supplement. The same thing is true for the **Knowledge Distributed Intelligence** (**KDI**) program. All proposals must be interdisciplinary. But that takes management. It is much easier to go in with one single little project and do your own thing. Cooperating can be very hard.

Judith Klavanz

ELSNews: The million dollar question: if you had a million dollars to spend on LR, what would you spend it on?

Klavans: I would spend it proportionally. Any resource project should have a certain set of components. The first component is the one we just mentioned, the collaboratory component: the first thing I would do would be to skim off 8%-10% towards travel for collaborative work, to make sure there are two or three different groups working on a particular resource. This makes it much easier to ensure that you don't go into some limited or parochial mark-up or structure that might be tailored to someone's particular research needs. The second thing would be a certain percentage, closer to 15%, that goes into determining standards for the project, whether we are talking a corpus, a dictionary, a lexicon, a collocation, or a terminology bank. That's 25%. Then I would put at least 50%-60% into content collection. And the rest I would earmark for maintenance, for stretching it out over a 5-10 year period. So the bulk would go towards the collection of content, but the other components are really important as well.

Then you try to leverage the money with industry, of course, see if you can double it...

FOR INFORMATION
Judith Klavans (klavans@cs.columbia.edu, http://www.cs.columbia.edu/~klavans/home.html) is Director of the **Center for Research on Information Access** at Columbia University, New York. Her research lies in computational linguistics and natural language processing. Prior to arriving at Columbia, she spent nearly ten years at the **TJ Watson IBM Research Division.** .

Projects

The **MATE** project aims to facilitate reuse of language resources by addressing the problems of creating, acquiring, and maintaining language corpora. The problems are addressed along two lines:

- through the development of a standard for annotating resources; and
- through the provision of tools which will make the processes of knowledge acquisition and extraction more efficient.

Specifically, MATE will treat spoken dialogue corpora at multiple levels, focusing on prosody, (morpho-)syntax, coreference, dialogue acts, and communicative difficulties, as well as inter-level interaction. The results of the project will be of particular benefit to developers of spoken language dialogue systems, but they will also be directly useful for other applications of language engineering. The project has recently started and is currently reviewing the state-of-the-art in dialogue corpora annotation and annotation toolkits.
The initial meeting of the project took place in Edinburgh on 23-24 April. There was a further Software Design Workshop in Edinburgh on the 15-16th June, where the basic design of the annotation tools was finalised.

MATE website: http://mate.mip.ou.dk

DISC aims to draw upon European experience in spoken language dialogue system development to produce a detailed, integrated set of development and evaluation methods and procedures. To do so, DISC is studying the current practice in development and evaluation of six aspects of state-of-the-art spoken language dialogue systems, identifying effective practices and deficiencies. The aspects concern speech recognition, speech synthesis, language understanding and generation, dialogue management, human factors, and system integration.

An analysis of 26 exemplars has been carried out according to a specified Grid and Life Cycle. The grid is used to characterise the system/component, and the Life Cycle attempts to capture the development process of the system/component. This analysis was summarised in six internal reports which have been made available to the **DISC Advisory Panel** (DAP). The project, started in June 1997, has just held the first DAP workshop and completed its first year review.
The next step will be to propose best practice guidelines and support tools which will be made available for testing.

DISC website: http://www.elsnet.org/disc

The aim of the ELSE project is to draw a blueprint for an evaluation protocol built around the paradigm of semi-automatic quantitative black box/grey box evaluation. This protocol is intended for both spoken and written NLP systems in the multilingual context of Europe.
The project started in January 1998 and will last 16 months. Initial activities have focused on drawing a picture of existing work in the domain of NLP systems evaluation, and refining the objective of the project. They resulted in a draft list of 30 potential evaluation tasks (very likely to be extended) that could be used for future campaigns, along with a very approximate estimation of the resources required to perform them. This result was presented at the LREC pre-conference workshop **Towards an open European Evaluation Infrastructure for NL and Speech**.
The next step will be to select five or six tasks from the above list on the basis of their greater interest for the field of NLP processing, and to draw up a detailed sketch of the infrastructure required to undertake them. Among the issues which are still unresolved are the multilingual aspect (in particular for resource re-use); how to evaluate systems involving dynamic adaptation, such as dialogues; and, to a lesser extent, how to articulate technology-oriented evaluation with, on one end,evaluation oriented towards users and applications; and on the other end, scientific and programme advances evaluation.

ELSE website: http://www.limsi.fr/TLP/ELSE/

The Corpor(e)al Infrastructure

*Mimo Caenepeel in conversation with **Uli Heid**.*

ELSNews: What do you see as the main achievements in the area of LR to date, and what do you consider the main technological challenges that are still outstanding?

Heid: I would say tools for robust parsing are one major achievement in the field of enabling technology. But they are still a challenge as well, because we are just seeing the first robust parsers for a few languages, and we are far away from having such tools for many languages. Another major achievement is that we finally have seed lexicons for several languages (as produced in the **PAROLE** project, for example). Building these lexicons is important because it helps get an even coverage of several languages.

As far as challenges go, on the one hand of course there is the challenge of having tools and resources for smaller languages. And the other thing is that we are still only half-way through exercises like parsing and identification of more meaningful structural components of text, which you can use for information extraction and similar applications. There has been a lot of progress in this area, in the last few years; but if you want to go for the identification of predicate-argument structures, or, in the medium term, for more semantically-oriented extraction, more is needed. And the next real medium-to-long-term challenge seems to me to be lexical-semantic description, and relating what we are currently doing for syntax to a lexical semantic description - and in the long and very long term, to sentence and discourse semantics.

ELSNews: German is probably one of the most resourced languages, after English. What has been going on in the development of resources for German? How is the effort organized, what is the role of universities, publishing houses and so on?

Heid: I'm not sure I would agree that German is one of the most resourced languages. It is in the sense that you have English far ahead, then three times nothing, and then a few languages like German, Italian and French... But German does not come close to what we have for English. Look at **WORDNET**, which has been prototypically developed for English and is now being extended to other languages in the **EuroWordNet** project. A German WORDNET is upcoming: there is a project in Germany devoted to this, and German is one of the languages of the EuroWordNet extension. German is resourced in the sense that there has been a tradition in IT companies and among builders of MT systems to build lexicons for their applications, but only some of these are available to R & D at large. German does not have a national corpus project so far. There is research going on in universities, people are dealing with material collected in an opportunistic way; but there is not yet a concerted action for resource building at a national level.

As far as lexicons go, Germany is not yet up to full speed with a country like the UK either: British publishers have a tradition of providing their dictionaries to universities and other institutions for research purposes (think of **Longman**, **Collins**, **Oxford**, most recently **Cambridge** ...). The main players in Germany are those universities which have a tradition in lexical and corpus research. And the **Institut für Deutsche Sprache** in **Mannheim**, which has become more active in the last few years in resource building, and is still accelerating its activities. Publishers are now starting to get into this type of activities. In Stuttgart we are cooperating with **Langenscheidt** publishers on the use of corpora to produce raw material for updating an existing commercial dictionary, especially with collocations.

An interesting trend in the field is the recent expansion in the development of corpora for smaller languages. Look at **Den Danske Ordbog**, the Danish dictionary project based on a major corpus. This is a Danish institution — halfway state-funded, halfway-private — building corpora and lexicons. Look at activities in the Netherlands, like the **CLVV** (**Commissie voor lexicale vertaalvoorzieningen**), a Dutch-Flemish organisation building bilingual dictionaries for Dutch, which as its first activity built a fully computerized lexicon for Dutch, the **Referentiebestand Nederlands**. Similarly, the Czech Republic has started, a few years ago, a major corpus and dictionary project. So, don't you think it's high time for Germany to move?

ELSNews: How do you use rather general resources like text corpora to build more specialized resources with added value, like lexicons? Is knowledge extraction possible, given that building lexicons by hand is time-consuming? And how would you say lexicons built (semi-)automatically compare with handcrafted ones built by lexicographers?

Heid: I see corpora as really infrastructural to lexicon building, and in the medium term also to grammar building. We know comparatively little about the more specific properties of lexical items. We have a good understanding of things like subcategorisation, for example verb complementation; but when it comes to things like which intensifier adverbs you typically combine with adjectives, or which adjectives you can use adverbially as modifiers of adjectives, nobody can provide this information off the shelf. You can, however, extract quite a bit of this from existing large-scale corpora, with reasonable results. So I do think that information extraction is possible. It's not always simple, and as we're seeing at this conference [LREC] the aquisition of lexical information from corpora becoming an NLP or resource-building technique in its own right. There are different approaches, statistical ones, symbolic ones, and hybrid ones which combine symbolic preselection and statistical procedures. I think it is a very important field, and one which will help us to create lexical resources in a reasonable time frame, and with comparatively reasonable effort. Definitely less effort than handcrafting them.

As an additional point there is validation, of two types. On the one hand you have the typical scenario of a linguist or a lexicographer trying to figure out the properties of lexical items for a dictionary. Such a person will basically produce what they are thinking of, and they might well leave out things which you would find in a general corpus. In such cases the corpus can correct and help balance the information which goes into the dictionary. For instance, it makes it possible to attach frequency strings not just to single words, but also to syntactic constructions, to collocations, and to certain (morpho)syntactic usages of words - the type of information we did not yet have in published dictionaries.

That's one side of validation. The other side is that manual work can help to counterbalance some of the problems we still have with information extraction from corpora. How much manual validation of the extraction results is needed, really depends on the intended applications. In the case of applications which rely more on statistical information, for example, you might well acquire information and just use what you've got: you might have some noise in there, but it won't matter, because it will not be of statistical relevance. On the other hand, if you are acquiring resources for symbolic systems — such as parsing within formal grammar, or machine translation — then it is important to still have some human intervention after the extraction. So you start from a corpus, you have a number of extraction routines, the outcome is candidate material to go into the dictionary, and then you have a human looking at that candidate material and basically trying to eliminate any noise which has slipped through the extraction machinery.

Usually, corpus-based dictionaries are only based on material of a certain type, like newspaper material. That means the extraction will miss out a number of uses, especially the ones with lower frequency — which lexicographers would have a tradition of looking at. So that is why lexicographers have always insisted on the need for what they call 'balanced' or 'broadly covering' corpora. We don't have this for many languages. We have it for English with the **British National Corpus** (BNC); and we are getting it for Danish and Czech ...

But we have discovered some interesting things while working for **Langenscheidt** publishers. We are good, on the basis of our news corpora, at certain types of lexical information which the press will typically need when talking about sports, economy, politics and so on. But we are bad at daily use, general items. And this is why I would definitely not want dictionaries created on the basis of opportunistically assembled corpora to replace our current hand-crafted dictionaries. But you can get more precise descriptions, you can get frequency information, collocations, and so on. And you discover things in the process. We are currently looking at adverbs combining with comparatives and superlatives in German. Many things are obvious, but some are real findings. And that, it seems to me, is computational linguistics in the most literal sense of the word: using computing machinery for linguistic discovery.

FOR INFORMATION
Ulrich Heid (heid@ims.uni-stuttgart.de, http://www.ims.uni-stuttgart.de/~uli/) works at the **Institut für Maschinelle Sprachverarbeitung**, Universität Stuttgart. He is a member of the ELSNET Executive Board. His research interests are in the areas of computational lexicography, corpus exploration and machine translation.

For more information on Stuttgart's cooperation with **Langenscheidt**, see **Docherty & Heid**: *Computational Metalexicography in Practice- Corpus-based support for the revision of a commercial dictionary*, in the Proceedings of the upcoming **Euralex-98** conference.
For more on

- **Den Danske Ordbog**: http://coco.ihi.ku.dk/~ddo/ddo_d.htm
- **CLVV** (Nederlandse Taalunie): clvv@ntu.nl

Language Resources and Evaluation: 10 Articles

1. At this moment, language resources are one indispensable key to unlock the potential of the global Information Society.
2. All sectors of society, and all languages, have an interest in seeing these resources developed, for a variety of purposes, economic, social, industrial and cultural.
3. Like human languages themselves, such resources are necessarily large-scale, and require a wide range of participants.
4. Although they are essential to realize the growth of private enterprise, they will not, indeed cannot, emerge simply from the sum of individual projects.
5. For each language, there is a need for strategy to co-ordinate existing resources and create new ones.
6. When resources have been created, there is a continuing requirement for support and maintenance.
7. These efforts for each language will benefit by taking into account, and profiting from, progress made in providing resources to underpin others.
8. Understanding of the role, usefulness and optimum means of preparation for language resources is a research theme in itself.
9. This co-operative understanding will benefit greatly from the use of common standards for evaluation of resources.
10. Cooperation can take many forms.

Nicholas Ostler unveils the 10 articles at LREC.

FOR INFORMATION
The full version of the 10 Articles is available from Nicholas Ostler (nostler@chibcha.demon.co.uk, http://www.bris.ac.uk/Depts/Philosophy/CTLL/FEL/).

Language Resources in Central and Eastern Europe

Tomaz Erjavec, Jozef Stefan Institute, Ljubljana

While Language Resources (LRs) for CEE languages are, in general, less developed than those for EU languages, recent years have seen a marked upsurge in available and publicised CEE resources. In many cases this is due to EU projects. The **Copernicus** Programme in particular not only provided funding for resource-oriented projects (e.g. MULTEXT-East, **Onomastica**) but also, either directly or indirectly (through **Awareness Seminars**, **ELSNet goes East**), raised awareness of their importance in CEE, not least among the funding bodies in these countries.

Of course, these language resources have not been developed from scratch in the last few years; but the recent focus on LRs in language technology has meant that they have been standardised, publicised and, crucially, made more widely available. **ELRA**, for instance, is starting to offer CEE resources in addition to the EU-language ones. Of particular importance to CEE LRs has been the Copernicus Concerted Action **Trans-European Language Resources Infrastructure** (TELRI). TELRI has connected CEE language technology centres with each other and with EU centres; it has produced a double CD-ROM, containing multilingual LRs of almost all CEE languages; and it has initiated the **TRACTOR** resource collection. **TELRI(-II)**, which will concentrate on the TRACTOR initiative, is to run for another three years. TELRI has now been established as a permanent association based in Germany, to maintain the action in the longer term.

As previously in the EU, the maturity of language technologies is influencing resource development in CEE. In Slovenia, for example, a publishing house recently went ahead — without government funding — with a project to collect a large reference corpus of Slovene; they now feel it is indispensable for producing quality dictionaries.

Finally, CEE LRs are being produced outside of their 'home countries' as well; at the LREC conference, for example, there was a presentation of an on-line corpus of Bosnian texts [1] from the University of Oslo, while a US project at CLR, New Mexico, has the Serbo-Croatian language included in its multilingual onomasticon [2].

It is difficult to give an overview of the kinds of resources that exist for CEE languages, because the situation differs so much from country to country. But in what follows I will give a general outline. This outline excludes Russian, which because of its very large number of speakers and its specific history has a special status among CEE languages. A great number of LRs have been produced in Russia; but unfortunately many are now being irretrievably lost, as there is no funding to maintain them, or the organisations that created them no longer exist.

In general, significant corpora have been or are being produced for a number of CEE languages (e.g. Romanian, Hungarian), often TEI-annotated and PoS-tagged. In many cases, recent EU funding helped with corpus projects, such as the Bulgarian corpora and resource tools produced by the **Linguistic Modelling Laboratory** of the Bulgarian Academy of Sciences. The largest to date is the **Czech National Corpus**, which currently contains almost a 100 million words; a large tree-bank is also being annotated for Czech. Both these efforts are funded by the Grant Agency of the Czech Republic. Some corpora are freely available, or even have on-line querying. On the whole, however, freely available or PoS-tagged corpora of the CEC languages are still scarce, and treebanks, large parallel corpora and sense-tagged corpora non-existent.

Estonia deserves to be mentioned in the context of machine readable dictionaries: in a happy marriage of (Soros) funding, copyright holders and language technology experts, they offer free WWW searches on a number of their dictionaries. However, this kind of availability of machine readable dictionaries is an exception rather than the rule. On the other hand, there is a growing number of (usually morphological) lexica available; for example, Bulgarian lexica are already being offered by ELRA, and lexica for six CEE languages have been produced by the MULTEXT-East project, and made available on the TELRI CD-ROM.

Speech processing is quite well-developed in a number of CEE countries. Speech resources have only recently become the focus of attention, often via EU projects. But now, due to the growing interest of large industries (like **Siemens**), speech corpora for a variety of settings (studio, telephone line), purposes (basic phonetic research, speech recognition, speech synthesis), and languages (Polish, Slovak) are being produced.

Lest the above sound too optimistic, it should be remembered that CEE LR development lags significantly behind EU languages. Quite a few CEE languages do not have their equivalent of the **Brown** corpus, for example. One reason for this is that government funding in CEE countries tends to be scarce, and EU funds insufficient. Moreover, the language industries have a harder time developing in these countries, and multinational/multilingual industries invest less in them. At LREC this was demonstrated quite well by the chart presented by **Microsoft** representative **David Brooks**, which showed the four bands in which **Microsoft** prioritises European languages for localisation. The first category was English, the second EU languages; 'major' CEE languages, i.e. those with a sufficient number of speakers/GNP, came third; and the fourth category had the 'minor' CEE languages. It is probably up to the EU to balance these categories with financial as well as political support.

References

Diana Santos: *Providing Access to Language Resources through the WorldWideWeb: the Oslo Corpus of Bosnian Texts.* In: LREC proceedings, Granada1998.

Svetlana Sheremetyeva, Jim Cowie, Sergei Nirenburg and Remi Zajac: *Multilingual Onomasticon as a Multipurpose NLP Resource.* In: LREC Proceedings, Granada1998.

FOR INFORMATION
Tomaz Erjavec (Tomaz.Erjavec@ijs.si, http://nl.ijs.si/tomaz | Jamova 39) is a researcher at the Department for Intelligent Systems, Jozef Stefan Institute, Ljubljana, Slovenia. His research includes work on linguistic resources, computational morphology, language technologies for the Slovene language, and typed feature-structure formalisms and implementations.

For more information about **TELRI**, see http://www.ids-mannheim.de/telri/
For a survey on Russian resources, see http://infomag.mipt.rssi.ru:8080/sections/lingvo.html
For more on Bulgarian corpora and resource tools: http://www.lml.acad.bg/

Reading Comprehension and Evaluation

Lynette Hirschman, The MITRE Corporation

A recurrent theme at LREC was the role of evaluation to both assess the state-of-the-art and to challenge researchers to focus on the next critical set of research problems. The use of standardized reading comprehension tests presents an ideal "grand challenge evaluation" for the language research community, for the following reasons:

- These tests would focus language research on a much richer notion of language processing, that goes beyond extraction towards *understanding*— not just *learning to read*, but *reading to learn* and retaining information from the material. This would encourage researchers to address the critical issues of machine language learning, including an interactive partnership (*reading together*) between the person supplying world knowledge and the computer integrating this into its vast store of data.
- The ability of computer systems to successfully pass such tests would provide built-in comparison to human capabilities. We can imagine a public relations coup for human language technology, akin to the success of IBM's Deep Blue chess program — when the first computer system graduates from elementary school, for example.
- Standardized tests are available for listening (speech input) as well as for reading, and cover foreign language learning.
- Reading comprehension would exercise emerging resources such as lexicons, WordNets, taxonomies, and knowledge bases, providing an indirect evaluation of these resources.
- Use of such tests for evaluation would reduce the cost of evaluation, by using "found" test material.

The paper I presented at LREC showed several sample tests, beginning with a test for beginning readers that turns out to be far too difficult for machines (see *Figure 1*). The reason for this is that eginning readers do not read very well, so it is easier to test them in terms of making proper associations between simple pictures and descriptive text. But for machines this level is too hard, because machine vision and scene analysis is not yet capable of doing this kind of analysis.

Fortunately, as children get older, the reading tests become more self-contained. They involve reading a story and answering questions based on information contained in the story. Indeed, one interesting kind of test asks who/what/when/where/why questions about the story, where the answers consist of phrases extracted directly from the text itself (*Figure 2*). This makes this task a domain-independent extension of current work on information extraction focused on a critical information access task, namely ability to answer ad hoc queries about stories.

At MITRE, we are actively investigating the feasibility of using standardized reading comprehension for evaluation of natural language technology. We are working to identify a first round of training material and blind test material, while also developing "**Deep Read**" — a natural language processing system to take the test. We invite any group interested in collaboration in this effort to contact us for further information.

A. The birds are on the flower.
B. The butterfly is on the tree.
C. The butterfly is near the flower.
D. The butterfly is under the flower.

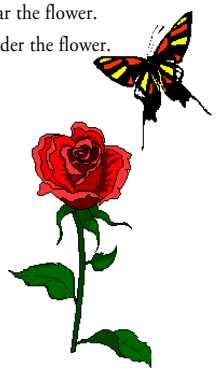


Figure 1: Sample Question 6-year-old level

How Maple Syrup is Made

Maple syrup comes from sugar maple trees. At one time, maple syrup was used to make sugar. This is why the tree is called a "sugar" maple tree.
Sugar maple trees make sap. Farmers collect the sap. The best time to collect sap is in February and March. The nights must be cold and the days warm.
The farmer drills a few small holes in each tree. He puts a spout in each hole. Then he hangs a bucket on the end of each spout. The bucket has a cover to keep rain and snow out. The sap drips into the bucket. About 10 gallons of sap come from each hole.

1. Who collects maple sap? **Farmers**
2. What does the farmer hang from a spout? **A bucket**
3. When is sap collected? **February and March/ cold days & warm nights**
4. Where does the farmer drill holes? **In the trees/in the maple trees**
5. Why is the bucket covered? **To keep out rain and snow**

Figure 2: "5 Ws" Sample Test

FOR INFORMATION
Lynette Hirschman (lynette@mitre.org, http://www.mitre.org/resources/centers/advanced_info/g04h/people.html#lynette) is the head of the **Intelligent Information Access Section** at MITRE. She is an active participant in MITRE's Natural Language research, which includes Alembic, Alembic Workbench, and Information Retrieval.

Evaluating Evaluation: US vs EU

Europe and the US have different approaches to funding in the area of Evaluation. In the DARPA approach, different groups are funded to do the same tasks; in EC initiatives, groups typically take on projects complementary to other ones. In the US model, there is a push to improve existing technology; the European model tends to be more user-driven. Should Europe move more towards the American model? And should it prioritize languages at policy level? We quizzed ***Joseph Mariani.***

ELSNews: *There are advantages to both the European and the American approach to funding. Do you feel that the EC should move more towards the American model?*

Mariani: The two programmes have different origins. The US programme started in 1984 (the same time ESPRIT started in Europe), and the idea was from the very beginning to look at the use of evaluation for accompanying research and checking the advances of a particular type of technology, to see whether it was worth the investment. This was very good for the technology — for European technology as well, because it could be tested in the American framework and shown to be of high quality. And it gave a clear picture of the state-of-the-art in various aspects of language engineering systems. Now that we have this picture, it could be that the US will turn more to application-oriented aspects (without abandoning the evaluation paradigm), since the technology may now be good enough for large sets of applications.

In Europe the approach was different. Here there is no competition between comparable laboratories in terms of technologies or of systems; there is competition in order to get a grant, but once the grant is obtained there is cooperation within each of the projects. Because of this, the technology evaluation aspect is missing in Europe. And as a result some application-oriented projects have probably used technology which was not good enough for the application which was targeted, and to some extent possibly wasted money. This is a problem, and we are probably just beginning to reverse this situation. The European approach was good for trying to find out which applications were of interest for the economy or society, the economy and so on. But being able to check the state of the technology and its progress is important too.

My hope for the future would be to have both aspects taken into account, both in the States and in Europe. Having a set of laboratories working on the same problem is a nice way of doing things: it makes it possible to compare and discuss results on the same basis, because the same data was used for conducting the test in an objective way. But once the technology is shown to work, you also have to check how well it works for a given application. We need both.

ELSNews*: LR are often referred to as a kind of infrastructure, like highways. On such a view, they should be publicly funded, and the general aim would be to build as many as possible. But another way of looking at it is that we need to know what LR are used for before we decide how to design them and how much money to spend. That would mean tailoring them to particular purposes and particular languages, and possibly prioritizing certain languages. What are your views on this?*

Mariani: In the computer age, languages which are not automatically processed are at a serious disadvantage. And for a language to be automatically processed you need a huge amount of data, both spoken and written, to develop and test systems and to study the language. So the availability of LR and the possibility to automatise a language go together. In this respect I think it is necessary to have LR for all languages.

On the other hand, the EC cannot cover LR for all languages, the effort is simply too large. The different linguistic communities and the different countries need to be involved too. In my view the role of the Commission is to initiate the process, set up the infrastructure, support the establishment of standards, enable distribution, and deal with other issues of issues of general issues like property rights. Then this general scheme can be enlarged by joint efforts to cover as many European languages as possible. Good communication between the EC and the member states is essential in this respect.

Then there are data which are of interest in terms of developing and training systems. The American **Broadcast News** task is an example of this: the data gathered for that can be used to train systems to recognise radio or TV broadcasting. That kind of resource requires a very large effort: you need to collect spoken data, and transcribe it in order to build an acoustic model for speech recognition; and then you need to collect an even larger set of data, and transcribe it into text, in order to train language models for recognition. Only the States have been able to provide data in that kind of quantity so far, and only for American English. Such an effort could not possibly be carried out for all European languages in one step. My view is that it is important to make a start: find the important application areas where we should develop such data, choose a set of languages, and proceed from there.

I would also like to stress the relationship between Evaluation and LR. You can develop LR and distribute them without caring about what will be done with them. But if you distribute

LR to a set of laboratories for the evaluation of systems, the data will have to be of high quality, as the different labs which will use the data will be able to check their validity and provide feedback. And that is one of the nice aspects of Evaluation: the need to set up a protocol, a fixed and well-organised methodology, with deadlines and schedules, and well-defined content. All this will contribute in turn to better LR. At the same time, the availability of LR which have been used for evaluation allows those laboratories which participated in the evaluation campaign to measure the progress achieved; and it enables those who didn't to compare their results with the state-of-the-art.

As I already said, launching an evaluation campaign is a very large effort. You cannot do it for 100 different tasks, 100 different systems, 100 different languages, so you have to make a choice. And you cannot use the same approach to technology evaluation and user-oriented evaluation. For general technology evaluation, you need to make a large effort on a generic task which is of sufficient interest to a large enough amount of laboratories. Whereas user-oriented evaluation is more specific: you take a particular application, and you test on it the quality of one system (or a small set of systems) specifically developed for that particular application. Both approaches are different, but complementary.

FOR INFORMATION
Joseph Mariani (mariani@limsi.fr, http:\\www.limsi.fr) is Director of **LIMSI-CNRS** (Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur, Orsay, France) and a member of the ELSNET Executive Board. His research interests are in the area of Spoken Language and Human-Machine Communication.

Future Events

Jul 13-24, 1998: *Robustness: Real Life Applications in Language and Speech, ELSNET's 6th European Summer School on Language and Speech Communication*, Barcelona, Spain. Further info: Email: summer98@gps.tsc.upc.es URL: http://gps-tsc.upc.es/veu/ess98/

Aug 01-04, 1998: *Discourse, Anaphora and Reference Resolution 2*, Lancaster University, United Kingdom. Further info: Email: eiamme@msmail.lancaster.ac.uk

Aug 4-8, 1998: *Euralex'98 International Congress , University of Liège, Belgium.* Further info: Email: Thierry.Fontenelle@sdt.cec.be URL: http://engdep1.philo.ulg.ac.be/euralex.htm

Aug 5-7, 1998: *Natural Language Generation*, 1998 International Workshop , Ontario, Canada. Further info: Email: hovy@isi.edu URL: http://logos.uwaterloo.ca/~inlg98

Aug 10-14, 1998: *COLING-ACL'98,* Montreal, Quebec, Canada. Further info: Email: coling-acl98-student@mpce.mq.edu.au URL: http://www.mri.mq.edu.au/conf/coling-acl98-student/

Aug 14-16, 1998: *Conference on Formal Grammar, HPSG and Categorial Grammar 1998* , Saarbruecken, Germany. Further info: Email: gj@ufal.ms.mff.cuni.cz URL: http://www.dfki.de/events/hpsg98/hpsg98-mailform.html

Aug 17-19, 1998: *MIND III: Irish Conference on Cognitive Science (Spatial Cognition)* , University College Dublin, Ireland. Further info: Email: hegarty@psych.ucsb.edu

Aug 17-21, 1998: *ESSLLI-98*, Saarbrueken, Germany. Further info: Email: adp@cs.city.ac.uk URL: http://www.cs.city.ac.uk/~adp/esslli98.html

Aug 21, 1998: *Third Australian Document Computing Symposium*, University of Sydney, Australia. Further info: Email: judy@staff.cs.usyd.edu.au URL: http://www.cmis.csiro.au/conferences-seminars/adcs98/

Aug 22-27, 1998: *8th Int. Conference on Human-Computer Interaction*, Munich, Germany. Further info: Email: HCI99@iao.fhg.de URL: http://hci99.iao.fhg.de

Aug 23-26, 1998: *Workshop on Text, speech and Dialogue (TSD'98)* ,Brno, Czech Republic. Further info: Email: kopecek@fi.muni.cz URL:http://www.fi.muni.cz/tsd98/

Aug 23-28, 1998: *ECAI-98*, Brighton, United Kingdom. Further info: Email: Henri.Prade@irit.fr URL: http://www.cogs.susx.ac.uk/call.html

Aug 24-28, 1998: *Coordination Technologies for Information Systems* (CTIS'98) , Vienna, Austria. Further info: Email: george@turing.cs.ucy.ac.cy URL: http://www.ifs.tuwien.ac.at/dexa98/

Sep 2-4, 1998: *SENSEVAL WSD Evaluation Exercise*, Sussex, United Kingdom. Further info: Email: senseval-coord@itri.bton.ac.uk URL: http://www.itri.brighton.ac.uk/events/senseval/cfp.txt

Sep 25-28, 1998: *The Future Of The Humanities In The Digital Age*, Bergen, Norway. Further info: Email: yvonne.bonete@ifi.uib.no URL: http://www.futurehum.uib.no/

Oct 26-29, 1998: *Speech and Computer (SPECOM'98)*, St-Petersburg, Russia. Further info: Email: specom@mail.iias.spb.su URL: http://www.spiiras.nw.ru/speech

ELSNET Participants

Academic Sites

NL	Utrecht University (coordinator)
A	OFAI/Univ. Vienna/Vienna Univ. of Technology
B	University of Antwerp
B	University of Leuven
BU	Bulgarian Acad. of Sciences, Sofia
BY	Belarussian Academy of Sciences, Minsk
CH	IDSIA, Lugano
CH	ISSCO, Geneva
CZ	Charles University, Prague
D	Univ. des Saarlandes/DFKI, Saarbrücken
D	Univ. Hamburg
D	Univ. Kiel
D	Univ. of Stuttgart
D	Ruhr-Univ. Bochum
D	Univ. Erlangen
DK	Ctr for Sprogteknologie, Copenhagen
DK	Ctr for PersonKommunikation (CPK), Aalborg
DK	Odense University
E	Universidad de Granada
E	Univ. Politecnica de Catalonia/Univ. Autonoma de Barcelona
E	Univ. Politecnica de Madrid
E	Univ. Politecnica de Valencia
F	LIMSI-CNRS, Orsay
F	IRIT, Toulouse
F	Inst. de la Comm. Parlée, Grenoble
F	IRISA, Rennes
F	Laboratoire Parole et Langage-CNRS, Aix-en-Provence
F	CRIN, Nancy
GR	ILSP/NCSR "Demokritos", Athens
GR	Wire Communications Lab., Patras
H	Hungarian Acad. of Sciences, Budapest
H	Technical University, Budapest
I	Ist. di Linguistica Computazionale, Pisa
I	IRST, Trento
I	Fondazione Ugo Bordoni, Rome
IRL	University College Dublin
IRL	University of Dublin
IT	Institute of Mathematics and Informatics, Vilnius
N	University of Trondheim
NL	Stichting Spraaktechnologie, Utrecht
NL	Inst. for Perception Research, Eindhoven
NL	Leyden Univ.
NL	Catholic Univ. of Nijmegen
NL	TNO Human Factors Research Institute
NL	Univ. of Amsterdam
NL	Univ. of Tilburg
NL	Univ. of Twente
P	INESC/ILTEC/Univ. Nova de Lisboa
PL	Polish Academy of Sciences, Warsaw
RO	Research Inst. for Informatics, Bucharest
RU	Russian Academy of Sciences, Moscow
S	KTH, Stockholm
S	Univ. of Linköping
UK	Defence Research Agency, Malvern
UK	UMIST, Univ. of Manchester
UK	Univ. of Cambridge
UK	Univ. College London/School of Oriental and African Studies (SOAS)
UK	University of Edinburgh
UK	Univ. of Essex
UK	Univ. of Dundee
UK	Univ. of Leeds
UK	Univ. of Sheffield
UK	Univ. of Sunderland
UK	Univ. of Sussex
UK	Univ. of Ulster
UK	Univ. of York

Industrial Sites

B	Lernout & Hauspie Speech Products
D	aspect GmbH
D	Daimler-Benz AG
D	Electronic Publishing Partners GmbH
D	Grundig Professional Electronics GmbH
D	IBM Deutschland
D	Langenscheidt
D	Novotech GmbH
D	pc-plus Computing
D	Philips Research Laboratories
D	Siemens AG
D	Verlag Moritz Diesterweg
DK	Tele Denmark
E	Telefonica I&D
F	ACSYS
F	Aerospatiale
F	GSI-ERLI
F	LINGA s.a.r.l.
F	MemoData
F	Rank Xerox Research Center
F	Systran SA
F	TGID
F	VECSYS Speech Processing
GR	Knowledge A.E.
H	Morphologic
I	CSELT
I	Database Informatica
I	Sogei (IRI-FINSIEL Group)
I	Syntax Sistemi Software
I	Tecnopolis CSATA Novus Ortus
I	Olivetti Ricerca SpA
NL	KPN Research Laboratories
NL	Polydoc N.V.
NL	University of Twente
RU	Analit, Ltd.
RU	Russicon Company
S	Telia Promotor (Call Centre Division)
FIN	Nokia Research Center
FIN	Kielikone Ltd
UK	ALPNET UK, Ltd
UK	BICC plc
UK	British Telecommunications
UK	Cambridge Algorithmica Ltd.
UK	Canon Research Centre Europe Ltd.
UK	Ensigma Ltd.
UK	Hewlett-Packard Labs
UK	Logica Cambridge Ltd.
UK	Sharp Laboratories
UK	SRI International
UK	Vocalis Ltd.

What is ELSNET?

ELSNET, the European Network in Language and Speech, was established in 1991 with funding from ESPRIT Basic Research. There were 25 founding members of the network. Currently, there are more than 60 universities and research institutes, and more than 45 companies participating.

The long-term technological goal which unites the members of ELSNET is to build integrated multilingual NL and speech systems with unrestricted coverage of both spoken and written language. Building multilingual NL and speech systems requires a massive joint effort by two pairs of communities: on the one hand, the natural language and speech communities, and on the other, academia and industry. Both pairs of communities are traditionally separated by wide gaps.

It is ELSNET's objective to provide a platform which bridges both gaps, and to ensure that all parties are provided with optimal conditions for fruitful collaboration. To achieve this, ELSNET has established an infrastructure for sharing knowledge, resources, problems, and solutions by offering (information) services and facilities, and by organising events which serve academia and industry in both the language and speech communities. In this respect, it is important to note that a network like ELSNET can only function well if all members of the network are prepared to give and to receive.

Electronic Mailing List

elsnet-list is ELSNET's electronic mailing list. Email sent to elsnet-list@let.ruu.nl is received by all Managing, Associate and Industrial node coordinators of the Network, as well as other persons who have an interest in ELSNET's activities. This mailing list may be used to announce activities, post job openings, or discuss issues which are relevant to people in the European natural language and speech communities. To request additions/deletions/changes of address in the mailing list, send mail to elsnet@let.ruu.nl.

ELSNET web pages

Detailed information about ELSNET and its activities and publications is available on the Web at the following URL: `http://www.elsnet.org`. Comments and suggestions for new web pages are very welcome.

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@let.ruu.nl
WWW: http://www.elsnet.org