elsnews 12.3

The Newsletter of the European Network in Human Language Technologies Autumn 2003

Lille hosts ELSNET's 11th Summer School

Computer Assisted Language Learning (CALL)

Latifa Al-Sulaiti, *University of Leeds*

Every year the ELSNET summer programme attracts a number of intellectuals, from eminent lecturers to experienced researchers and novice postgraduates, to focus on a specific topic in the field of language and speech communication. Just as there is a change of topic each year there is also a change of location. This gives participants the dual benefit of learning both intellectually and culturally. This year, 35 participants from universities in the Netherlands, Spain, France, Greece, UK, Denmark, Italy, Romania, Sweden, Belgium, Ireland, and from the Basque region all came to Lille University to attend the summer programme. Among the participants were some non-European students from USA, Algeria, Syria, UAE, and Qatar. The beautiful warm weather we had gave us the chance to enjoy our stay in Lille and its surroundings.

The two-week programme was intensive with two-hour sessions in the mornings and two-hour sessions in the afternoons. Students' presentations were also incorporated into the programme. It is difficult, therefore, to review in detail all the topics that were covered.

Participants at the ELSNET Summer School in Lille

Inside this issue:

Autumn 2003

elsnet

ISSN 1350-990X

However, in these comments I will mainly touch on sessions that related to my work and interests in the design of a corpus of contemporary Arabic for TAFL (Teaching Arabic as a Foreign Language).

The programme began with lectures from Wolfgang Menzel, of the University of Hamburg, on *Error diagnosis and feedback generation for language tutoring systems*. His lectures focussed on illustrating two approaches to error diagnosis: error anticipation and anticipation-free methods. He explored the application of these methods in the domains of syntax and morphology, pointing out their merits and demerits. He also suggested better ways of obtaining feedback by using methods such as graphics rather than giving explanations for the errors. I found some parts of his lectures to be hard to follow, especially as the course progressed. On the whole, however, I benefited from some ideas, especially those that related to designing exercises for learning Arabic. Although the course was technical it was eased by the good sense of humour of the lecturer.

The second contribution to the programme featured lectures by Jozef Colpaert of the University of Antwerp. His lectures were entitled *Introduction to CALL*. He gave a very useful and comprehensive introduction to CALL, covering its history, research, tools, design, tutors, engineering, and much more. His lectures were enjoyable and easy to follow for most students, especially as they were enlivened with demonstrations of some CALL applications and by his witty character. I found this course to be very beneficial to my own interests as, in addition to absorbing new ideas, I obtained a number of useful resources to explore and learn more from, such as websites, books, and articles.

In the afternoon there were parallel practical sessions. One focussed on language and the other on speech. In the language session, which was directed by Wolfgang Menzel, students were helped to develop mini web-based tutoring demos. Although most students found it too technical, as it required knowledge of the programming language Prolog, some persistent students managed to do the task even with no programming knowledge.

Michael McTear, of the University of Ulster, headed the speech sessions, which covered the use of the CSLU toolkit for designing and implementing spoken dialogue systems. The toolkit was user-friendly and great fun to use. For this reason the majority of students chose to do it. I personally liked the idea of developing dialogues for teaching a language and I enjoyed the sessions. Unfortunately, however, the program does not work for Arabic.

One bad thing about the parallel sessions was that it was not possible to try both, as some students wished. Nevertheless, a few managed to have a taste of each.

There were other sessions by Philippe Martin of the University of Paris 7 and the enthusiastic Aline Germain-Rutherford of the University of Ottawa, who introduced us to the software 'WinPitch' and showed us its versatility. The feature I found unique about this software was having the pitch of the target language appear on the screen for the learner to emulate and compare his pitch with the target pitch. Another attractive feature is the fact that the teacher can give visual feedback to students. It was a pity, though, that the time was too limited to go through everything satisfactorily. Furthermore, the number of students was high and conditions in the lab uncomfortable. Although this course was entitled *Introduction to Speech for CALL*, it was more like a demonstration of how 'WinPitch' software works and how it can be used for language learning rather than an introduction to speech.

During the second week we were occupied with lectures presented by Lars Borin of the University of Gothenburg, on *Introduction to Natural Language Processing (NLP)*, and Anne-Marie Oster, of KTH Stockholm, on *Systems for pronunciation training*.

Anne-Marie discussed and gave interesting demonstrations of programs used for teaching children with special needs (hearing problems). Although teaching for special needs is not my area I found the lectures, and especially the demonstrations, quite inspiring, especially if the methods could be adapted for teaching Arabic as a foreign language. It would be especially useful as Arabic contains a number of difficult sounds.

Of particular interest to my research were the lectures given by Lars Borin. He stressed the use of authentic text (corpora) in language learning. In his practical sessions he introduced us to some language applications such as the 'Compleat Lexical Tutor' and the 'Glosser'. The former contains several concordance programs for research, teaching, and tutorials. Learners can choose a corpus among the available list and conduct their language search. As for the latter, it is a program that facilitates reading by providing users with an online dictionary and morphological analyser. It was mainly developed to teach Dutch speakers French but can be used for English as well. We also explored some programs for tagging, chunking, and lemmatising. However, none of the applications can be used for Arabic. I must say that the time devoted to NLP tools was too limited to get a grasp of how they work. In addition, some programs ran on Linux and not many students knew how to use it.

The students' presentations were divided into groups based on their field. Although they touched interesting

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

Editor: Lynne Cahill.

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

ISSN 1350-990X

FOR INFORMATION
Contributions to *ELSNews*, and corrections, should be sent to:
lynneca@sussex.ac.uk

Tel: +44 1273 678659
Fax: +44 1273 671320

Material for the next issue is due:
12th December 2003

and varied topics that were worth reviewing, it is impractical to go through them all in this short review. However, they are available on the Summer School website (see below).

The programme's social events were busy and wonderful. They were a mixture of cocktail parties and visits to important landmarks in and around Lille, such as the Museum of Arts and the old city of Lille. There was also the opportunity to see the museum of Modern Art in Villeneuve d'Ascq and attend the celebration of Bastille Day. The Excursion to Boulogne-sur-Mer and its old town brought a lively variety to the programme and made participants more at ease with each other. Students planned their own weekend programmes, some heading to nearby exquisite destinations such as Paris and Bruges. The programme ended with a lovely dinner at the Alliance Hotel where we had a taste of rabbit paté, which was a relief to some students who were expecting to be served 'les escargots'! The farewell barbecue party was the highlight of the social activities. Students, teachers, and organisers all came together in a friendly and enjoyable atmosphere.

Even though the programme had some negative points, such as the limited access to computers, not having the reading lists ahead of time, and the uncomfortable conditions in the lab, meeting students from different cultural and scientific background and teachers with high expertise made it a worthwhile summer activity.

FOR INFORMATION

Latifa Al-Sulaiti is a research student in the School of Computing at the University of Leeds

Email: Latifa@comp.leeds.ac.uk

Web: www.compleeds.ac.uk

Summer School web site:
www.univ-lille3.fr/ess2003

contd from p. 7

tion, and it has been taken up by many companies working in speech applications. The VoiceXML Forum stages the SpeechTEK show every year to showcase the latest technologies in text-to-speech and other voice technologies. This year the 373 companies belonging to the forum demonstrated a wide range of products with VoiceXML capabilities.

The Swedish company PipeBeach, whose Chief Technical Officer (CTO) is VoiceXML forum leader Scott McGlashan, has recently been acquired by Hewlett Packard after the two companies worked together on a voice portal for a French telecommunications company. HP wants the VoiceXML expertise of PipeBeach to layer on top of their OpenCall media platform. McGlashan will become CTO of HP's OpenCall unit.

Three significant applications were presented at SpeechTEK from AT&T, ScanSoft, and VoiceGenie. AT&T launched its VoiceTone service in limited areas, hosted through 21 internet sites in the US. This service allows users to interact with computer-based systems with little or no human involvement. ScanSoft showcased its SpeechWorks platform, which features speaker verification, text-to-speech in 22 languages, and speech recognition in 46 languages. The VoiceGenie platform, recently upgraded to version 6, now includes toolsets for managing and configuring elements on the platform. This illustrates the progress being made towards the use of VoiceXML in complex speech systems.

The future

Further development of VoiceXML 2.0 is still under way, with many requested changes still to be incorporated. Other changes suggested by developers and other interested parties have been deferred until after 2.0 is finalised. Issues relating to call control, lexicons, natural language semantics, and embedding in a multimodal context are also under consideration and specific proposals will be developed by other committees within the Voice Browser Working Group as well as other W3C Working Groups.

FOR INFORMATION

[1] This quote and information in this article taken from articles and press releases on the PipeBeach and VoiceXML Forum web sites.

VoiceXML Forum:
Web: www.voicexml.org

The VoiceXML Forum welcomes suggestions. Requests for further changes to the language, comments on the language or the development process can be submitted to www-voice@w3.org.

Voice Browser Working Group:
Web: www.w3.org/Voice

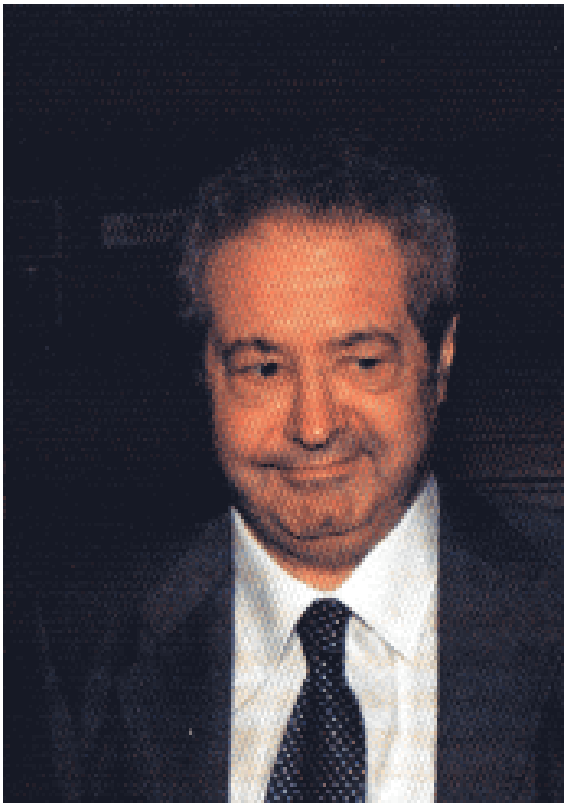
PipeBeach/HP:
Web: www.pipebeach.com
Email: scott.mcglashan@hp.com

Obituary

Antonio Zampolli: a remembrance

Maghi King *ISSCO/TIM/ETI, University of Geneva.*

In August Antonio Zampolli died suddenly in Pisa. As one of the most influential people in the Computational Linguistics community in Europe, his loss will be greatly felt. First, ***Maghi King*** *gives a personal perspective. Below* ***Steven Krauwer*** *remembers his role in the founding of ELSNET.*



Antonio Zampolli (Picture courtesy of ILC)

As did so many others, I learned of Antonio Zampolli's death by fire with an enormous sense of shock. Zampolli was part of the permanent intellectual landscape of my life. As a very naïve young lecturer in computation, I came across computational linguistics almost by accident. The idea of combining computing (my then present) and the study of language (my then past) was very attractive, and I set about finding out more. My quest led me to go to Coling in Pisa in 1973. That trip was a turning point. Everything contributed to making it magical. I had never been out of the UK before, I travelled in a beaten up old banger driven by a graduate student scarcely younger than myself, it was a great adventure, and even the weather was beautiful. And then we arrived in Pisa; soaked in antiquity and sunshine, alive with political strife, architecturally astonishing – small wonder that I was ripe to fall in love. And that I did as soon as I went to the first sessions of the conference. There was so much passionate debate, so much friendly talk. The world was smaller then, and even though everybody seemed to know everybody else, I never felt left out. And at the centre of all this was Antonio Zampolli. He seemed to be everywhere at once, to be involved in all the debates, in all the social life of excursions and banquets, and in all the sheer fun of it all. He was the ringmaster of a circus I desperately wanted a part in.

I returned to Pisa the following year for one of the Pisa Summer Schools, and the experience was confirmed. The courses were given by people I had only dreamed of meeting, discussion amongst the participants, staff and students alike, was all-encompassing and enthralling, lasting often into the early hours of the morning. And once again, this was something created by Antonio Zampolli, who played as hard as he worked, was passionate about his chosen field, and ready to do all he could to advance it.

Paradoxically, I saw less of Zampolli once the move to ISSCO had brought me geographically closer to him. Although I am sure that I must have seen him in the intervening years, my next really clear memory of him is in 1978, in Luxembourg, the evening before the meeting which was to give the first concrete impetus to the creation of the Eurotra project. By chance I was staying in the same hotel as Zampolli. When I went down to dinner in the hotel dining room, he was already there, at a table with Bernard Vauquois and others whose identity I no longer remember. Nobody recognised me, and I watched with a kind of envy as they talked, laughed, discussed intensely – and made a great deal of noise. Later I was to be present, as so many of us were, at dinner parties made sparkling by his grace and his wit.

Once the preparation of Eurotra started, I saw a great deal of Zampolli: he was present in most of the meetings of the first two or three years, and came often to meetings of CETIL, the committee to which we reported. What I remember most from those meetings is the quality of his mind. He was not the sort of person who studied all the documents in meticulous detail before a meeting, and he could never stay in a meeting for long without feeling the sudden need to make a telephone call or to visit a Commission official in his office. Nonetheless, his role was major. He could spot a weakness in an argument before most of us had started to absorb its first steps, just as he could make valid

and valuable connections in what looked like wild leaps of intuition, although I suspect that what was really behind them was a depth of knowledge and a width of experience uncommon in his contemporaries and even rarer in his juniors. In short, he had one of the best analytic minds it has ever been my privilege to come across.

He was also a very adroit politician. If I say that part of his political talent was to be able to spot a gap in the academic market and work to fill it, I hope that that will not be construed too cynically. When he said in the early 1990s that the greatest need in human language technology at the time was to develop resources and standards for resources, he was right, and almost every European research worker in the field has benefited in one way or another from the insight. It was the same with the Language Resources and Evaluation Conferences (LREC) that Antonio Zampolli launched in 1998. I went to the first pleased that finally there was some public recognition of evaluation as a discipline in its own right and of the importance of developing linguistic resources, but expecting to find a rather small number of like minded enthusiasts. Many of us were surprised and delighted at the large number of participants and at the enthusiasm that once again reigned around Antonio: yet again, he had got something right.

If I had to choose just three words to describe Antonio Zampolli, they would be energy, warmth, and intelligence. The world is poorer for his leaving of it: we will remember him.

FOR INFORMATION

Maghi King is Director of ISSCO, Geneva.

Email: Margaret.King@issco.unige.ch

Web: www.issco.unige.ch/staff/king/king.html

Remembering ELSNET's founder

Steven Krauwer, *ELSNET*

Antonio Zampolli was one of the founding fathers of ELSNET back in 1991, and he was the driving force behind numerous initiatives in the world of language and speech technology. Those who have worked closely with him will remember him as a creative scientist with a broad interest and a very strategic mind, and as a good friend and colleague.

He was one of the people who have really changed the world of language and speech technology, not by inventing some new and clever syntactic parser or translation tool, but rather by his remarkable capability to detect and exploit opportunities, possible synergies, and future directions for our field long before most of us had even started thinking about them.

He knew many people and many people knew him, not just in Europe, but in the whole world, and he used his contacts to bring people together to take new initiatives.

He was a member of the ELSNET Board from the very beginning in 1991 until his sudden death. I met him first in the early seventies when I had just joined the field, and later on many times during the Eurotra machine translation project, where he was a member of one of the supervisory committees.

When I took over as coordinator of ELSNET from Ewan Klein in 1995, our cooperation became more intense and our meetings more frequent, and I was always amazed to see how full of ideas he was. During meetings he could look as if he was dozing away, and then suddenly he opened his eyes and sketched us an appealing, new vision, idea, or initiative that would change our world. I looked forward to every meeting with him, and until his health started causing him problems some two years ago he never missed a single meeting we organised. And if the meeting took place in Pisa or elsewhere in Italy he always managed to organise wonderful excursions (including excellent meals) for those of us who had bought APEX tickets and had to stay over for an extra day at the weekend.

Life will really be different without him, and we will all miss him, not only as an inspiring colleague, but also as a very dear friend – and as a great and charming entertainer at our working dinners!

FOR INFORMATION

More tributes to Antonio Zampolli can be found via the ELSNET web site: www.elsnet.org

Also, the Text Encoding Initiative has a web page for tributes: www.tei-c.org/Publicity/AZ/

Feature

VoiceXML: transforming the telephone industry

Lynne Cahill, *ELSNews*

In the present climate of standardisation and markup languages, it may seem that the speech and language communities are more distant than ever. The typical markup languages used by NLP practitioners around the world, SGML, HTML, and especially now XML are all primarily intended for text applications. Speech technology, on the other hand, is creeping gradually into the business world, with speech synthesis used in many telephone situations around the world. The emergence of the VoiceXML language as a widely accepted standard in speech applications promises to bring the two areas together, providing a 'markup' language for speech.

What is VoiceXML?

VoiceXML (or VXML), the Voice Extensible Markup Language, is a markup language for building interactive speech applications. Based on XML, it provides a high-level programming interface to speech and telephony resources for application developers, service providers, and equipment manufacturers. The language has been designed to give application developers full control over spoken dialogue between the user and the application.

Documents in VoiceXML can describe: spoken prompts; output of audio files and streams; recognition of spoken words and phrases; recognition of touch tone key presses; recording of spoken input; control of dialogue flow; and telephone control (e.g. hangup).

The procedure is as follows. When a user makes a call to a voice portal the VoiceXML platform sends a request to a URL for a VoiceXML application. The dialogue proceeds by the user speaking, the application converting the speech (and any other input such as touch tone presses) into VoiceXML using automatic speech recognition. The system's responses are presented to the user using text-to-speech synthesis, audio files, or streaming media.

A session begins when the user starts to interact with a VoiceXML interpreter and continues as different VoiceXML documents are loaded. The session ends when the user hangs up, or when requested by the user. The user can interrupt the system at any time to request another service.

Uses and applications

VoiceXML is intended to improve the customer's experience while reducing operational costs. Instead of talking to a human, the computer language allows the user's requests to be translated into a form the computer can manipulate. The computer application can then find the relevant answer in the company's database and use text-to-speech technology to construct an answer. Using VoiceXML, a company is able to take the "yes", "no", "maybe", and even the "fine" and "yeah" comments and put them into a database, dramatically streamlining the telephone support process. It is the first step towards full automation of the telephone enquiry process.

According to Dave Raggett, W3C voice browser activity leader, VoiceXML is not only of interest to the business sector as it could open up a whole new world to people who might not otherwise be able to use the internet. "People will be able to interact via spoken commands and listening to synthetic speech and music," he said. "This will also benefit people with visual impairments or needing Web access while keeping their hands and eyes free for other things."[1]

Some examples

Like any other markup language, VoiceXML breaks down what the human wants into something the computer can work with. The voice browser working group gives the following example. A customer calls their local pizza delivery shop and says "I would like a medium coca cola and a large pizza with pepperoni and mushrooms". A possible representation for this after speech recognition might be:

```
{
drink: {
beverage: "coke"
drinksize: "medium"}
pizza: {
pizzasize: "large" topping: ["pepperoni",
"mushrooms"]}}
```

The typical VoiceXML scenario (Picture courtesy of PipeBeach/HP)

This example shows a high-level interpretation of the speech input. An example including lower-level speech features can be demonstrated with the following specifications of a system's output (i.e., synthesised speech):

```
<form>
<field name="travellers">
<prompt>How many are travelling?</prompt>
<grammar mode="voice"
        src==www.example.com/number.grxml
        type=="application/grammar+xml">
</field>
</form>
```

This example shows how close to XML the language is. The grammar element here references an external speech grammar, *number.grxml* and specifies that the prompt should be realised using "voice". More detail again can be seen in the following example, specifying how the prompt should be realised:

```
<form>
<field name= ="travellers">
<prompt>How <emphasis> many </emphasis>
are travelling?</prompt>
</field>
</form>
```

This example makes use of the SSML (Speech Synthesis Markup Language). Elements from SSML, such as the 'emphasis' here, can only appear inside the prompt element. VoiceXML 1.0 provided its own speech markup elements, but these have been dropped from release 2.0 in favour of using the existing SSML elements.

History

The first release of VoiceXML was published by AT&T, Lucent, IBM, and Motorola in 2000, a development of the Phone Markup Language project begun by AT&T in 1995. Following a World Wide Web Consortium (W3C) conference on voice browsers in 1998, these companies created the VoiceXML Forum to pool their efforts and define a standard.

The development of version 2.0 from 1.0 has involved three areas. It was agreed that 1.0 offered approximately the right level of functionality but required improvements in the areas of interoperability, functional completeness, and clarity. A process of consultation has been in train throughout the development, with forum members and any other interested parties requested to submit suggestions for improvements.

Although there are similarities, VoiceXML is a more comprehensive approach to the question of voice applications than the SALT (Speech Application Language Tags) specification which has also been submitted to the W3C. The latter is a set of lightweight extensions to existing markup languages. While VoiceXML concentrates on telephony applications, SALT focusses on multimodal speech applications.

Current status

The Voice Browser Working Group of W3C has adopted VoiceXML 2.0 as a proposed recommenda-

contd on p. 3

Handling errors in spoken dialogue

Michael McTear, *University of Ulster*

Handling errors is one of the most challenging tasks for developers of spoken dialogue systems. Should error-handling be based on how humans handle errors in human-human interaction? How might dialogue systems detect errors and recover gracefully from them? How do users behave when faced with errors produced by a spoken dialogue system?

These, and many other questions, were the subject of a recent ISCA tutorial and research workshop. For many participants one of the most interesting discussions focussed on the seemingly wide gap between academic research on spoken dialogue and the concerns of commercial developers of dialogue systems to be deployed in the real world. One of the invited speakers, Mike Phillips from Scansoft, discussed user interface design for spoken dialogue systems from a commercial perspective, while Herb Clark of Stanford University illustrated timing issues in dialogue, based on detailed analyses of human-human communication.

Scansoft is involved in the deployment of a large number of telephone-based dialogue systems, some of which handle more than 100,000 calls per day. Telephone-based dialogue is recognised as a major issue for speech recognition. Yet, as Mike Phillips pointed out, in-grammar speech recognition errors tend not to be the major source of error. Rather the problems lie with false acceptance of out-of-grammar utterances and with mismatches between applications and users' expectations. To deal with these errors typically a system-directed dialogue is recommended with fairly constrained methods for handling errors – for example, using explicit confirmations. A lot of attention is paid to the careful design of prompts that should reduce the potential for error. Fine-tuning of the deployed system results in various changes, of which almost 50% involve expanding the grammar. Summing up progress so far, Phillips concluded that we are still a long way from dialogue systems that can produce behaviour similar to human-human interactions.

How humans behave in dialogue was the topic of Herb Clark's presentation. In particular, Clark illustrated how people not only choose what to say but also when to say it. Many so-called errors are complex phenomena that demonstrate how speakers are constrained by issues of timing when they talk. One example of timing involves processing constraints, as speakers cannot speak until they have formulated what they are going to say. However, speakers also attend to other types of timing

Cross timing requires that speakers place their utterances with respect to their partner's speech (turn-taking). Speakers are also constrained by internal timing, which involves trying to deliver their utterances fluently. Departure from internal timing constraints often signals specific information. For example, fillers (such as "um" and "er"), which are usually factored out in speech recognition grammars as 'noise', fulfil the important function of signalling a delay in the production of a fluent utterance that carries the implication that the speaker does not wish to relinquish the turn. Many more examples of such timing phenomena were illustrated in Clark's paper.

Initially it seemed that these two perspectives on handling spoken dialogue errors were on different wavelengths. What could detailed analyses of human-human dialogue have to say to designers of commercial spoken dialogue systems? On reflection, however, it seems that there is much that is of use, possibly even in the short-term. As Clark showed, humans communicate a lot of information, not only in what they say but also in how they say it. System logs of dialogues contain crucial data in the form of prosodic and paralinguistic cues as well as features associated with timing and this information could be used by designers of spoken dialogue systems to assist with detecting when things are going wrong in the dialogue. Many of the other papers presented at the workshop illustrated in detail how issues such as these are being addressed in current approaches to spoken dialogue technology.

Michael McTear

FOR INFORMATION

Michael McTear is Professor in the Faculty of Engineering, University of Ulster. His book "Spoken Dialogue Technology" is due to be published by Springer Verlag.

Email: mf.mctear@ulster.ac.uk

Web: www.infc.ulster.ac.uk/staff/mf.mctear

Workshop web site: www.speech.kth.se/error/

STO: A lexical database of Danish for language technology applications

Anna Braasch, *Center for Sprogteknologi, Copenhagen*

The Centre for Language Technology (Center for SprogTeknologi, CST) is running a national project with the aim of developing a large size Danish lexicon for natural language processing, including commercial language technology products and computational linguistic research purposes. The short name for the project is STO, which stands for SprogTeknologisk Ordbase in Danish (i.e., Lexical Database for Language Technology). CST is a Danish government research institute under the Ministry of Science, Technology and Innovation. Its mission is to carry out and promote strategic research and commercial development in the fields of language technology and computational linguistics. The project gets funding from the ministry for a period of three years, finishing at the end of February 2004.

Project objectives

The objective of the STO project is the development of a computational lexicon of Danish for a broad practical application area. Language industry applications and research into computational linguistics often suffer from the lack of a large-scale comprehensive lexicon. This appears to be a bottleneck for most applications. In particular, for less widely spoken languages such as Danish it is essential to develop some multipurpose and flexible language resources in order to optimise the cost/benefit ratio.
In this sense the lexicon being developed in STO will serve as a basic lexical data collection from which various dedicated lexicon modules can be derived for particular applications, such as lemmatisers, inflection analysers and generators, shallow parsers, or Danish modules for MT systems.

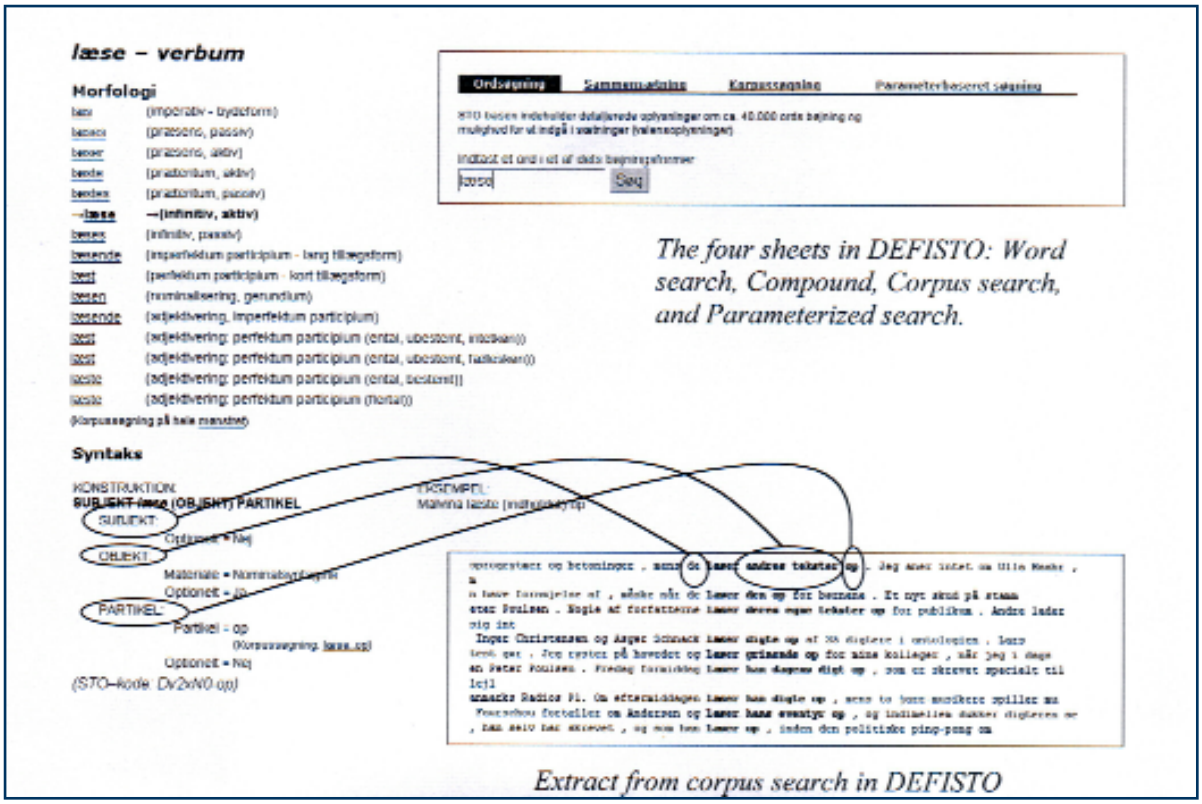
Anna Braasch

Current project structure

CST started the project and is responsible for the project management and the co-ordination of work. Further, various central tasks such as software development, elaboration of linguistic specifications, and setting up guidelines for encoding the lexicon are carried out at CST. The co-operating project members are affiliated to three different institutes, one being located at the University of Copenhagen, another at the Copenhagen Business School, and a third at the University of Southern Denmark. The project members (16 part-timers in number) have rather different professional skills within the area such as theoretical linguistics, terminology, lexicology, corpus linguistics, computational lexicography and linguistics, and database knowledge.

Background and project tasks

The Danish STO project greatly benefits from the experience acquired in the multi-lingual LE-PAROLE (1996-98) and LE-SIMPLE (1998-2000) projects, as regards the model, descriptive language and methodology of linguistic description. Also the lexicon material, which was developed at CST within the framework of these projects, is integrated into the STO lexicon. Although this material provided us with a well-established basis, considerable work had to be carried out in addition to the enlargement of the vocabulary. Main tasks were elaboration of comprehensive linguistic specifications, tailoring the format and structure of lexicon entries, establishing firm guidelines for lemma selection and encoding, compiling corpora of domains for language for specific purposes (LSP), etc. Also a number of reusable tools were developed for corpus lemmatisation and effective encoding of linguistic features.



DEFISTO search example: the verb læse *(to read). DEFISTO presents all inflected forms and syntactic constructions. Only one of several constructions is shown here: "(...) they read other people's texts aloud" (Example prepared by Nicolai Hartvig Sørensen)*

Coverage of STO

The agreement with the ministry stipulated that the size of STO will be 45,000 entries fully described with morphological and syntactic information on all parts-of speech. Of these, 30,000 belong to the general language, mainly as used in newspapers. The rest (15,000) is made up of entries from six different domains of LSP – namely IT, health, administration, environment, commerce, and finance. The corpora have been collected from the web using an onomasiological structure of the domain in question for identification of relevant sites and searching relevant texts. The texts selected reflect present-day communications from experts to laymen. The selection of these domains and texts is based on their close relatedness to people's everyday life and the fact that the vocabulary covered is frequently used together with general language words. The objective is to ensure coherence in the coverage of the lexicon. It is assumed that also domain languages will be included in the coverage required in all-round LT applications to come.

During the project development we realised the need for a much larger number of general language words with exhaustive morphological description being especially useful for recognition purposes. We therefore increased this part of the lexicon substantially and it is now twice as large as originally planned.

Linguistic features in STO

The project is mainly concerned with the formalised representation of existing linguistic knowledge for computational use, e.g. the treatment of inflectional morphology, noun compounding, or syntactic constructions. In order to prepare the lexicon for a broad utilisation within natural language processing and related applications, we are keeping also an eye on upcoming national and international standards for language resources, especially lexicons. It is important to ensure to the highest possible degree that the Danish STO lexicon follows the tendencies of standardisation to facilitate future links with other lexicons in large multi-lingual lexical databases.

The linguistic content of the lexicon is organised into three independent but coherently linked layers, namely the morphological, syntactic, and semantic layer. The full linguistic description of a lemma is structured according to these three layers and it can be obtained as a whole, like a dictionary entry, by accessing the whole set of descriptive units linked to the lemma in question. In general, NLP applications demand explicit and very detailed linguistic descriptions; STO meets this demand by consisting of well-defined pieces and structures of information.

At the morphological level the main information given about a lemma concerns its part-of-speech, inflectional features, and for nouns compounding is included too.

All spelling and inflectional variants are included in order to ensure proper recognition. On the other hand, all variants are also provided with information about whether the variant is approved by the Official Danish Spelling Dictionary (Retskrivningsordbogen) or not, which ensures the generation of proper forms only. Of course, obvious misspellings, etc. are left out. As regards the syntactic behaviour of lemmas, we record all syntactic patterns based on evidence and frequency in the corpus. In STO, verbs have especially rich syntactic combinatory features where the possible syntactic functions and phrase structures governed by a verb are included. STO also accounts for the syntactic features of adverbs using the same description method as for adjectives, nouns, and verbs which is quite unusual in NLP lexicons. A PhD project of which the aim is to find a method to describe the lexical syntactic properties of Danish adverbs in a computational lexicon is running in parallel with STO, and the results are going to be implemented in the lexicon. This goes for both the general and domain language vocabulary, and this makes the STO lexicon a particularly valuable resource, since these kinds of information are normally not provided in traditional domain-specific language dictionaries.

In the STO lexicon the semantic description is provided at three different levels. The most detailed level of semantic description (level 3) corresponds to the information types provided in the SIMPLE project including ontological typing, domain information, semantic relations in terms of qualia structure, argument structure, event structure, and selectional restrictions.

As regards the domain languages, lemmas belonging to five of them are provided with basic domain information (level 1). In between these two levels, we defined a lean semantic description (level 2) for test purposes, providing information about sub-senses and ontological relationships of the lemma, in addition to the specification of the domain. The health domain serves as test vocabulary for level 2 semantics.

Application areas

The STO material is, in its current form, already being used in a number of projects and applications. Typical uses include a search engine for content based information retrieval (OntoQuery), a Treebank of Danish, a morphological generator for Danish nouns, a tagger, and a lemmatiser for Danish. An obvious way to test the material will be the integration of the lexicon into a machine translation system with Danish as one of the languages. Further application areas such as teaching Danish as foreign language (grammar and spelling checkers) are evident too.

Project status and perspectives

The STO project is now in its last phase. For people to who can read Danish, it is possible to get some idea of the material by visiting the project web site (see below). It is the first version of an internet-based user interface accessing a part of the STO lexicon. It is possible to browse and search for morphological and syntactic information, and also corpus evidence of the key word. An English version of the user guidelines is also under development.

Although the current version of the STO lexicon contains a large amount of data, there are still a number of relevant tasks to cope with, not only increasing the lexical coverage. An extension to indude pronunciation is desirable and would be feasible. We are faced with the challenge of semantics – the capturing of semantic features and sub-categorisation rules is one of the most important outstanding points on our list. On the methodology side, we aim at automating several processes to a higher degree, such as domain ontology building, lexical profiling from corpus data, and statistically based lemma selection procedures.

FOR INFORMATION

Project web site: www.cst.dk/sto

Contact:
Anna Braasch, Project Manager
Email: anna@cst.dk

Project members:
Center for Sprogteknologi
Nicolai Hartvig Sørensen
Lina Henriksen
Costanza Navaretta
Dorte Haltrup Hansen
Lene Offersgaard
Sussi Olsen
Bolette S. Pedersen
Claus Povlsen
Sanni Nimb (PhD research)

Copenhagen Business School,
Institute for Computational Linguistics
Web: www.id.cbs.dk
Stig W. Jørgensen, local Project Manager
Jette Drost
Carsten Hansen

University of Copenhagen
Institute for General and Applied Linguistics
Web: www.chpling.dk
Ole Nedergaard Thomsen
Mikkel Hald

University of Southern Denmark, Kolding
Henrik Holmboe
Email: holmboe@asb.dk

Opinion

On Resources and Standards

Annie Zaenen, *Xerox PARC*

Antonio Zampolli died on the 22nd of August. Among many other things, he was an energetic promoter of HLT (Human Language Technology) in Europe (and one of the founders of ELSNET) and for the last twenty years, a tireless advocate of the creation of European linguistic resources and standards. As such he was among the creators of EAGLES (Expert Advisory Group on Language Engineering Standards) and of ELRA/ELDA (the European Language Resources Association) as well as LREC.

The availability of language resources has profoundly changed the way computational linguistics is practised over the last 15 years and evidently, having clean text (or speech) in a variety of languages is an essential prerequisite for the competitiveness of European HLT and for research. One only wishes that ELDA's corpora list were longer.

Let me indulge in a personal peeve here. Often the distributors of resources make a distinction between private and academic institutions in the prices they set. For instance, some corpora distributed by ELDA have different prices for academic institutions and for private ones, even when the resources are used for research. Clearly, when resources are used for commercial purposes they should be treated as commercial goods but I don't see the rationale for the distinction when we are talking about research. I would like a corpus to study the use of auxiliaries in Dutch. If I were working for a university the PAROLE corpus would cost 300 Euros, now it would cost my lab 1300 Euros. This research has no commercial applications and it even has no bearing on the rest of what my lab does, but I get the reaction, "Your company can afford it." Maybe, but that is not the question: any company-dependent research unit has its own budget and has to live within its means. It is not obvious that a private lab is better off financially than an academic one and it is also not obvious these days that the academic institution is not engaged in commercial transactions. I know of quite a few cases where resources have been produced financed solely with the taxpayer's money and then given or licensed to private companies by a university lab at the discretion of the lab's director to compete with products that were created through private investment.

Annie Zaenen

But back to the importance of linguistic resources. We know how to produce clean text and speech, even if there will always be debates about the importance of balanced corpora. It is less clear how to produce useful resources that go beyond unannotated corpora. For instance much energy has been spent on trying to create reusable electronic lexicons with little effect. Lexicons are created with a theory of lexical organisation in mind and here we are not just talking about resources but also about standards. The European approach for a long time has been: let's get everybody together, let's agree upon a scheme and then get some funding to get the work done. A lot of energy goes into the agreeing stage and the funding is never sufficient to get the work done. In fact, it is fair to say that the only widely used lexical resource is the English WordNet. This was a resource that was not made by a committee. Everybody that uses it has something to critise and deplore but it exists and it is used. Versions for other languages were produced the European way. When I last looked they were all too small to be of much use.

Another widely used resource is the Penn Treebank. Again it was not created by committee initiative but it has profoundly changed the way syntactic development is done. Treebank efforts have been reported for nearly all European languages and for several others. During most of the 1990s a language could be defined as a dialect with a morphological analyser, now it is becoming a dialect with a Treebank.

The English Treebank seemed to have imposed a *de facto* standard. Most of us would not have voted for its encoding scheme if we had had a chance but within its own assumptions it is well done and, as Elisabet Engdahl and I argued years ago, syntactic therories differ more in notation than in substance so everybody has been able to adapt the Treebank to their needs. Notations are, however, not totally with-

out importance and the Treebank encoding has proved not to be the most convenient one. Treebanks for other languages (although not the Chinese Treebank that follows more the parochial East Coast theoretical fashion) don't use constituent structure style encodings but adopt the more European tradition of dependency grammars and the Penn Treebank community is now following this trend with its Propbank. In German (and other Germanic, verb-second languages) the treebank efforts have also been part of a movement within the computational community towards topological parsing schemes, again an older tradition that is felt to be more adapted to the language. We see then that even when the initial standard is 'wrong', it gets corrected by the Computational Linguistics community. Practical pressures seem to lead to the 'right theory'. If I interpret these trends rightly, they show that there is really no reason to agonise about *a priori* standards by committee before getting down to useful work: the requirements of use will lead to the correct standards much quicker. And if a resource doesn't get used it certainly doesn't matter whether it follows standards or not.

FOR INFORMATION

Annie Zaenen is Principal Research Scientist at Xerox PARC, USA

Email: zaenen@parc.com
Web: www2.parc.com/istl/members/zaenen

Contd from p. 14

Another social experience on the DiaBruck boat trip

The workshop started off with a tutorial on *Best Practice in Empirically-based Dialogue Research* given by Laurent Romary, Michael Strube, and David Traum, focussing on corpus analysis: why people do it, what use it is, how to do it, and (importantly) how not to do it. Materials from the tutorial can be downloaded from the conference website, as can copies of the twenty contributed papers and abstracts of the posters, demos, and invited talks, as well as Ivana's photo gallery.

The final event was a hike to the Teufelsberg or 'Devil's Castle'; your correspondent can tell you little about this as he was one of a sensible contingent who turned back when it started raining. Serendipity intervened, and we found ourselves at a village festival called the 'Powai-Feschd' in nearby Gisingen. We never found out what Powai meant but it appeared to be celebrated by sitting around drinking beer, eating sausages and waffles, and listening to cheesy 1980s pop music.

Next year's gathering of this multi-disciplinary community will be a new departure as it will be the first workshop in the series to be held in Southern Europe. Enric Vallduví will be welcoming participants to Barcelona, the capital of Catalonia. One of the hardest tasks for a conference chair is coming up with a name that indicates the location and topic of the event without sounding too ugly or comical. Enric has opted for a name (suggested by Massimo Poesio) which strikes this correspondent as particularly felicitous: Catalog.

FOR INFORMATION

Rodger Kibble is Lecturer in Computer Science in teh Department of Computing, Goldsmiths College, London.

Email: R.Kibble.gold.ac.uk

Web: igor.gold.ac.uk/~mas01rk/

Workshop web site: wcoli.uni-sb.de/diabruck/
Thanks to Ivana Kruijff-Korbayová for the pictures

Workshop report

DiaBruck 2003

Rodger Kibble, *Goldsmiths College, University of London*

DiaBruck 2003 was the seventh in a series of workshops on the semantics and pragmatics of dialogue that have been taking place annually since 1997. This year it was the turn of the University of the Saarland to host the event, which was ably chaired by the ubiquitous Ivana Kruijff-Korbayová.

Although the name 'DiaBruck' was composed from an amalgam of 'dialogue' and 'Saarbrücken', the event actually took place some distance from the university's home town in the tranquil setting of the Hotel Scheidberg in the countryside near Saarlouis. The relatively isolated location and relaxed atmosphere encouraged the type of informal contacts which are essential to the success of a conference: there were ample opportunities for networking and exchanging ideas over enormous German breakfasts, lavish lunches, and buffet dinners.

Fittingly, the conference turned out to be a dialogue of sorts among researchers from widely differing backgrounds, since just about every sub-discipline in the linguistic sciences (and many beyond) has some contribution to make to the study and simulation of human conversation. A few of the paradigms represented in the contributed papers were: experimental psychology, corpus analysis, philosophy of language, formal semantics, game theory, symbolic AI, and social theory. The invited speakers represented this diversity in microcosm, with backgrounds in formal semantics (Nicholas Asher), experimental psychology (Martin Pickering), and AI logics (Andreas Herzig).

The demo session attracts some interested parties

contd on p. 13

Undeterred by the rain, Diabruck participants at the Devil's Castle

Calendar

Future Events

Nov 20-21 *Translating and the Computer 25:* London, UK
Email: nadamides@aslib.com URL: www.aslib.com/conferences/#TC25

Nov 21 *Workshop on Balkan Language Resources and Tools:* Thessaloniki, Greece
Email: vangelis@iit.demokritos.gr URL: www.iit.demokritos.gr/skel/bci03_workshop

Dec 1-2 *Second International Workshop on Dictionary Writing Systems:* Brighton, UK
Email: adam@itri.bton.ac.uk URL: www.itri.bton.ac.uk/courses/dec03

Dec 8-9 *Fourth Dutch-Belgian Information Retrieval Workshop:* Amsterdam, Netherlands
Email: ir@science.uva.nl URL: lit.science.uva.nl/DIR

Dec 18 *Second CoLogNET-ELSNET Symposium:* Amsterdam, Netherlands
Email: Raffaella.Bernardi@unibz.it URL: www-uilots.let.uu.nl/~ctl/workshops/CES03

Dec 19 *14th Meeting of Computational Linguistics in the Netherlands (CLIN):* Antwerp, Belgium
Email: bart.decadt@pcger65.uia.ac.be URL: cnts.uia.ac.be/clin2003

Jan 6-7 *Seventh Annual CLUK Research Colloquium:* Birmingham, UK
Email: mgl@cs.bham.ac.uk URL: www.csbham.ac.uk/~mlg/cluk

Submission deadlines

Nov 15 *First International Jpint Conference on Natural Language Processing (IJC-NLP04):* Hainan Island, China
Mar 22-24, Email: tsujii@is.s.u-tokyo.ac.jp

Dec 1 *RIAO 2004 (Coupling approaches, coupling media and coupling languages for information retrieval):*
Avignon, France, Apr 26-28, URL: www.riao.org

Dec 15 *Coling 2004:* Geneva, Switzerland, Aug 22 (tutorial proposals),
Email: coling2004-tutorials@cui.unige.ch URL:www.issco.unige.ch/coling2004

Dec 15 *Coling 2004:* Geneva, Switzerland, Aug 22 (workshop proposals),
Email: hess@cl.unizh.ch, URL:www.issco.unige.ch/coling2004

Jan 15 *TALN'04 (Traitement Automatique du Langue Naturel):* Fez, Morocco, Apr 19-22
URL: www.lpl.univ-aix.fr/jep-taln04

Jan 15 *ADS'04 (Affective Dialogue Systems):* Kloster Irsee, Germany, Jun 14-16
URL: www.sigmedia.org/ads04

Jan 15 *SPECOM'2004 (Speech and the Computer):* St Petersburg, Russia, Sep 20-22,
URL: www.spiiras.nw.ru/speech

This is only a selection – see www.elsnet.org/cgi-bin/elsnet/events.pl for details of more events and deadlines.pl for more deadlines.

If you would like to write a review of any of these (or other language/speech related events you attend), please contact the *ELSNews* editor.

ELSNET

The ELSNET Participants: Academic Sites

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

Industrial Sites

B	Dhaxley Translations
CH	Localization Industry Standards Association
D	Novotech GmbH
D	Sympalog Speech Technologies AG
D	DaimlerChrysler AG
D	Langenscheidt KG
D	Verlag Moritz Diesterweg GmbH
D	aspect Gesellschaft für Mensch-Maschine Kommunikation mbH
D	Philips Research Laboratories
D	Grundig Professional Electronics GmbH
D	Acolada Gmbh
D	IBM Deutschland
D	Varetis Communications
D	Heartsome Europe GmbH
D	Xtramind Technologies GmbH
D	Sympalog Voice Solutions GmbH
D	Scansoft Aachen GmbH
DK	Tele Danmark
DK	Zacco A/S
E	SchlumbergerSema sae
E	Telefonica I & D
EL	KNOWLEGDE S.A.
F	LINGA s.a.r.l.
F	Systran SA
F	Xerox Research Centre Europe
F	Memodata
F	Aerospatiale
F	VECSYS
F	SCIPER
F	TGID
FIN	Kielikone Oy
FIN	Nokia Research Center
HU	MorphoLogic Ltd.
I	OLIVETTI RICERCA SCpA
I	LOQUENDO
LV	TILDE
NL	Compuleer
NL	Knowledge Concepts BV
NL	Sopheon NV
NL	IP Globalnet Nederland BV
PL	Neurosoft Sp. z o.o.
RU	Russicon Company
RU	ANALIT Ltd
S	Sema Infodata
S	Telia Promotor AB
UK	Vocalis, Ltd.
UK	Hewlett-Packard Laboratories
UK	Canon Research Centre Europe Ltd
UK	ALPNET UK Limited
UK	Sharp Laboratories of Europe Ltd
UK	BT Adastral Park
UK	Logica Cambridge Ltd.
UK	20/20 Speech LTD
UK	Tangent Telecom Ltd
UK	Cambridge Algorithmica Limited
UK	Fourth Person Ltd.

What is ELSNET?

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

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

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

ELSNET seeks to develop an environment which allows optimal exploitation of the available human and intellectual resources in order to advance the field. To this end, the Network has established an infrastructure for the sharing of knowledge, resources, problems, and solutions across the language and speech communities, and serving both academia and industry. It has developed various structures (committees, special interest groups), events (summer schools, workshops), and services (website, e-mail lists, *ELSNews*, information dissemination, knowledge brokerage).

Electronic Mailing List

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

Subscriptions

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

FOR INFORMATION
ELSNET
Utrecht Institute of Linguistics OTS, Utrecht University,
Trans 10, 3512 JK, Utrecht, The Netherlands
Tel: +31 30 253 6039
Fax: +31 30 253 6000
Email: elsnet@elsnet.org
Web: http://www.elsnet.org