

EUROMAP releases Human Language Technology Scorecard

Lynne Cahill, ELSNET

The final UK event of the EUROMAP Language Technologies project was held in London earlier this month, presenting the results of the 'mapping' aspects of the project to the community in the form of a scorecard.

The aim of the EUROMAP project was to improve the path-to-market for the Human Language Technologies in Europe by providing awareness, bridge-building and market-enabling services. One of the final deliverables of the project is the scorecard – a report outlining the current state-of-the-art in all of the European countries. This report focusses not only on the level of the research work in each country, but also on how well research and development has been transferred to the market.

The day was hosted by the UK's EUROMAP partner, ITRI (the Information Technology Research Institute of the University of Brighton) and was held in the CBI Conference Centre in central London. Around seventy delegates attended, mostly from industry. The day was kicked off by Donia Scott of ITRI, who gave a general overview of the EUROMAP project before handing over to Andrew Joscelyne of the EUROMAP Coordination Team, co-author of the full scorecard document which is to be published later this month. Joscelyne's presentation gave a fairly detailed overview of the findings of the project, including highlighting the strengths and weaknesses of each country in respect of its HLT.

The overview of the situation in the UK was presented by Nicholas Ostler. Standing in for Rose Lockwood, main author of the scorecard, who was unfortunately

unable to attend, Ostler pointed out that the UK is one of the leading three players in HLT in Europe and suggested that future progress will be dependent on extending the horizons to include other media and other (minority) languages.



Nicholas Ostler

After lunch, speakers from industry addressed issues of the market place. Ray Jackson of Solcara discussed the kind of technologies that will be needed in the future in a world driven by data, information, and knowledge. He particularly drew the audience's attention to the distinction between these three entities, and suggested an extension to the T.S. Eliot quotation "Where is the wisdom we have lost in knowledge? Where is the knowledge we have lost in information?" to include the question "Where is the information we have lost in data?" He made the point that universities and industry can work together profitably, illustrating it with a number of case studies.

David Horowitz of Vox Generation Ltd. described



Donia Scott, Roger Evans, and Amy Neale of ITRI

Inside this issue:

Clarity: Cross-lingual Information Retrieval

Mark Sanderson 3

Opinion

Annie Zaenen 6

Seville hosts multilingual workshop

Julio Gonzalo and Anselmo Peñas 8

New ELSNET members

9

Dialogue corpora from DARPA

John Aberdeen 11

Language in the Knowledge-based Society: an outlook from Romania

Dan Tufis 12

Calendar

15

Spring
2003

elsnet
.....

ELSNes 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 Krauer and Brigitte Burger.

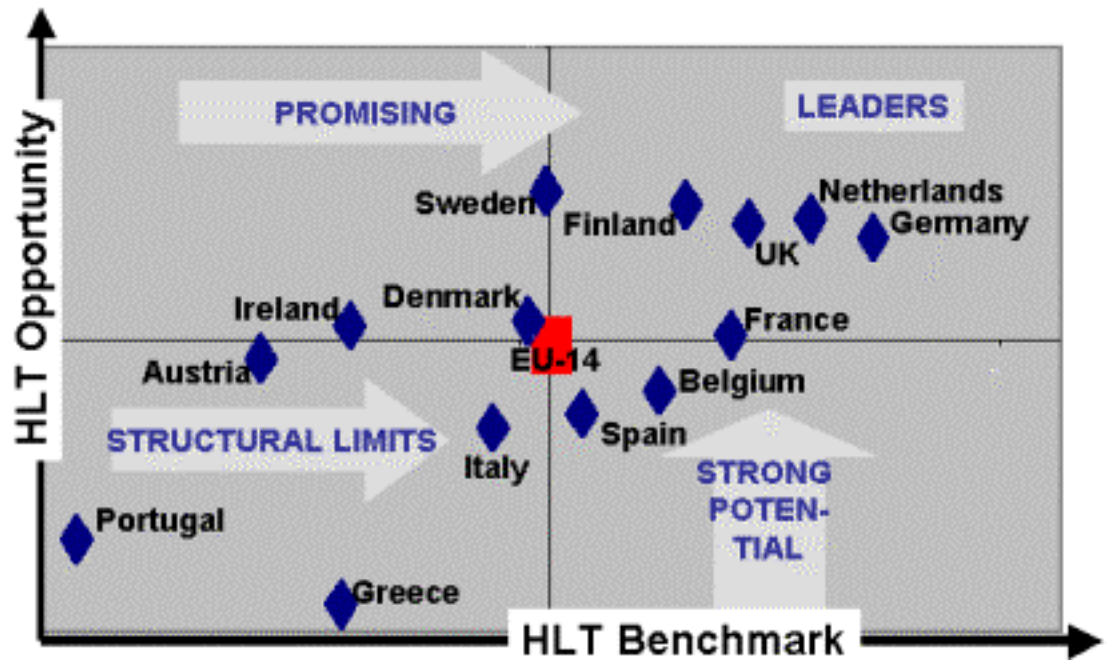
ISSN 1350-990X

© ELSNET 2003

FOR INFORMATION
Contributions to
ELSNes, and corrections,
should be sent to
lynneca@cogssusx.ac.uk

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

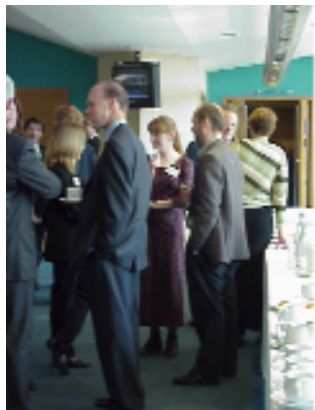
Material for the next
issue is due:
27th June 2003



The graphical version of the scorecard illustrates how the human language technologies in the EU countries are progressing, taking into account the opportunities each country has. Countries performing above expectation for their opportunities appear in the bottom right, those performing below appear in the top left corner. 'Opportunity' was determined by third party research considering factors such as ease of business formation, access to ICT etc. and weighted for relevance to HLT.

the FASiL consortium, funded by the EU, which involves a number of industry and academic partners and is developing a 'Flexible and Adaptive Multi-Modal Spoken Interface Language' with the aim of providing a new standard of conversational speech interaction. User groups involved with this project include the Royal National Institute for the Blind and the Royal National Institute for the Deaf.

Pete Walters, the UKISHelp coordinator, gave a very interesting and helpful presentation on the opportunities for funding in FP6. Although, as he pointed out, it is too late now to be starting to think about the first call, he highlighted areas of the second call that could be appropriate for HLT. He particularly stressed the benefits for SMEs of participation in EU-funded projects, making it very clear that, while the EU route is not appropriate for



Delegates enjoy a chat over coffee

companies with an idea that needs a quick path-to-market, for longer term R&D plans, EU funding can bring great benefits.

The day was rounded off by Andrew Joscelyne completing his overview of the state of the HLTs in Europe with a look at the future members, mostly in the East. He

highlighted the strengths and weaknesses of all of the prospective member countries and gave an optimistic outlook for HLT across Europe.

The main conclusion to be taken from this event were that the EUROMAP project has done a very useful and timely job of mapping the strengths and weaknesses of the Human Language Technologies in the various European countries. The full scorecard report will undoubtedly give much more detail about potential progress in each country. It is to be hoped that the bridge-building and information providing aspects of the project will be continued, possibly under the auspices of ELSNET, when the EUROMAP project winds down.

FOR INFORMATION

For EUROMAP contact Amy Neale:

Tel: +44 1273 642906

Fax: +44 1273 642908

Email: amy.neale@itri.bton.ac.uk

Web: www.hltcentral.org/euromap

Slides from the presentations at this event are available online. The final scorecard will also be available at the project website.

e

Clarity: Cross-lingual Information Retrieval

Feature

Mark Sanderson, University of Sheffield

This article describes the Clarity project: a 36-month EU 5th Framework RTD project. The project consortium consists of three academic groups and three companies. The project has run for just over two years and here we report on the goals of the project and the results gained so far.

Introduction

Cross-lingual information retrieval (CLIR) is the retrieval of documents written in one language with queries written in another. Such systems typically work as follows: queries written in one language (referred to as the *source language*) are automatically translated in some manner into the language of the document collection (the *target language*) and retrieval takes place. It is possible to translate the whole document collection instead of each query, but this approach is not often pursued. CLIR has been studied for around ten years. Over the decade, researchers have identified the major problems of translating queries for retrieval; determined a series of approaches to building and utilising a variety of

translation resources; and produced a number of retrieval algorithms that address the problems of CLIR ensuring a high quality ranking of documents resulting from a cross language search. It is a problem area where in a relatively short space of time, academia has made CLIR a workable, though not perfect, technology. Although a great deal of effort has gone into making the technology of CLIR work, particularly for languages for which a great many translation resources exist, very little research studying the users of CLIR systems has taken place, nor has there been much research examining CLIR for languages where few translation resources exist, so-called *low density languages*. The aim of Clarity is to explore such aspects of the subject.

In the rest of this article, usability testing and ways of dealing with low density languages are described followed by conclusions and details of the Clarity consortium.



clarity

search

Search in:

☒ English
 ☐ Finnish

For documents in:

☐ English
 ☒ Finnish

Date range:

last two weeks

Max documents:

10

racism

search

☐ see query translated

list

report

overview

sorted by rank, sort by language, sort by date

Norrbäck puhuu rasismia vastaan

☒ Norrbäck puhuu (make a speech, drawl, spoken) rasismia (racism) vastaan (against, versus, v., vs) (until, not till, only just, just now, a moment ago, bath whisk)(answer, reply, respond, react, answer for) #1, +, 2002.11.05, 68 words long, **search terms:** racism, **terms found:** racism
keywords: yleinen (general, prevalent, rife, widespread) vaan (but, on the contrary) tuomita (judgement, disapprove, give, award) lama (recession, slump, stagnation) kotimaa (native land, home country, native country, fatherland, original home)

save to Keep List

Rasisminvastainen junakierue kulkee halki Suomen

☒ Rasisminvastainen (racism)(opposite, facing, adverse, anti-, unfavourable, unfavorable, upward) junakierue kulkee (journey, travel, be carried along) halki (in two, in two parts, apart, across) #2, +, 2002.11.05, 185 words long, **search terms:** racism, **terms found:** racism
keywords: ilta (night) matka (voyage) varsova (Varsova, Warsaw)

save to Keep List

Figure 1: Prototype user interface displaying results of a query, sorted by rank. Each result contains a document summary and a number of keywords, displayed in both the target language and the source language translation.

Spring
2003

elsnet
.....

Feature (contd.)

Users and usability

When considering the users of a CLIR system, one must first tackle an important question. One of the first things that most people will ask when hearing about CLIR is, why would anyone want to retrieve documents they presumably cannot read? There are a number of answers:

- A searcher may not be able to read the target language, may have access to a human translator but may have too many documents to be translated. By retrieving documents for a particular query of interest and passing the top N only to the translator, they may make more efficient use of that person.
- Some searchers may read a foreign language but struggle to query in it. If a CLIR system can search accurately, then it could be a useful solution for such searchers.
- It has also been said that media companies that deliver their output in languages spoken by few people wish to use CLIR systems. Many of the potential audience of the media company's output will most likely be able to speak another more widely spoken language, often English. Such people are less likely to expect relevant information in their native language and when searching, only query in the more popular language. The media companies want to attract the audience back to their 'minority language content' through CLIR.
- Even people who can read and write many languages (polyglots) may want a CLIR system as they could enter a query once and have documents returned written in all the languages they know. Such people constitute the user groups in the Clarity consortium.

It was the needs of the last group that informed the design of the Clarity system: journalists and translators, from two of the companies in the consortium, are polyglots who wish to search in multiple languages, but would prefer to enter their query only once. A user-centred interface design methodology was adopted. By monitoring existing practices with monolingual search engines and discussing interface mock-ups at the two companies (BBC Monitoring and Alma Media), it was then possible to compile a list of user requirements. These requirements were subsequently used to create a prototype user interface. At present, the prototype has been tested at the two industrial consortium sites. In the final year of Clarity, the system will be extensively evaluated and user-tested on site with the user groups. Figure 1 shows the interface in its current state.

In the final version of the system, two additional interface features will enable users to better interact with a CLIR system:

- The first is a document-organisation tool based on concept hierarchies that show a dynamically generated hierarchy of words extracted from the retrieved documents arranged with general terms at the top leading to related more specific terms below. The terms can be shown in whichever language the user prefers.
- The second feature is a retrieval report produced after each search. Here key terms from documents and the styles of the documents retrieved (analysed automatically) are described.

The system has been implemented using a novel distributed architecture that allows the separate

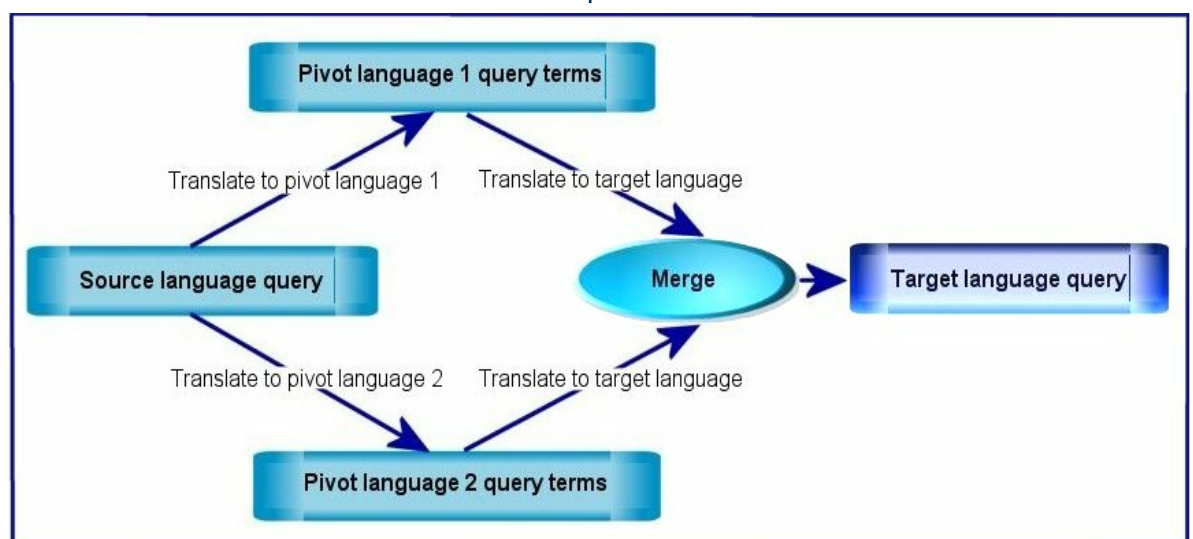


Figure 2: Lexical triangulation – by translating the original query into multiple pivot languages then into the target language, a set of query terms is produced. When these query terms are merged, this produces a more effective target language query than when a single pivot language is used.

e

development sites (University of Sheffield, University of Tampere, SICS, Tilde) to build their component parts locally and integrate them through an Internet based protocol. Communication between the components is through SOAP. This is proving to be a most valuable approach, as Clarity has been able to start building a prototype much faster and cheaper than would have been the case if the conventional approach of centralising integration had been taken.

Low density languages

Much existing CLIR research assumes the availability of translation resources that, for the majority of languages, do not exist. A second focus of Clarity is low density languages: languages for which there are few translation resources. Here two approaches to aiding CLIR are being investigated: first, using multiple bilingual dictionaries for transitive cross-language retrieval via one, possibly many, so-called pivot languages (see Figure 2); and second, using advanced n-gram techniques for translating words not in dictionaries (e.g., proper names). The two methods are now described.

Triangulated translation

Most approaches to CLIR assume that resources providing a direct translation between the query and document languages exist. Such an assumption, however, is often false. In such cases, an intermediate (or *pivot*) language provides a means of transitive translation of the query language to the document via the pivot, at the cost, however, of introducing many errors. Clarity has attempted a novel approach of translating in parallel across multiple intermediate languages and fusing the results (see Figure 2). Such a technique removes the errors, raising the effectiveness of the tested retrieval system, up to and possibly above the level expected, if a direct translation route had existed. Across a number of retrieval situations and combinations of languages, the approach has proved to be highly effective. One question that might reasonably be asked of the triangulation work is, do any CLIR users actually need or have the translation resources for triangulation via multiple pivot languages? It turns out that such examples exist. Good translation facilities from Latvian and Lithuanian into English exist, as do translation resources into Russian. However, no other translation resource for other languages currently exist online. For these languages triangulation is a necessity. At the half-way point in the Clarity project, it extended its scope and language coverage through the addition of a Latvian partner, Tilde: a Baltic language technology company. Now, the two Baltic

languages will, via triangulation, be added to the Clarity system.

Untranslatable words translated

A common problem when translating a user's query via some form of translation resource, for example, a bilingual dictionary is finding that some or all of the query words do not occur in the dictionary. Such a problem is particularly common with proper nouns: names of places, and in some languages, names of people are commonly spelled differently. Clarity is exploring the use of both n-gram and skip-gram techniques to locate likely candidate translations. A user enters a query, one of the words of which is not in a translation dictionary. Words in the document collection that appear to be close misspellings of the query word are chosen as possible candidate translations and added to the query. Even though the query becomes a combination of correct and incorrect translations, initial results testing on existing document test collections show the technique to be promising, indicating that Information Retrieval systems are tolerant of a level of error present in a query.

Conclusion

Cross language retrieval is a topic that has been long studied with clear progress made. However, certain key aspects of CLIR have not been studied – usability and low density languages – success in both is key if CLIR is to be adopted by those who wish to use it. Clarity is building a retrieval system based on such research. Results indicate the approaches taken are promising.

Feature (contd.)

FOR INFORMATION

Mark Sanderson is Lecturer in the Department of Information Studies at the University of Sheffield

Email: m.sanderson@shef.ac.uk

Web: disshef.ac.uk/mark

CLARITY:

Email: clarity@shef.ac.uk

Web: www.dcsshef.ac.uk/research/groups/nlp/clarity

The CLARITY Consortium comprises the following partners:

University of Sheffield, UK
University of Tampere, Finland
Stiftelsen Swedish Institute of Computer Science (SICS), Sweden
BBC Monitoring, UK
Alma Media, Finland
Tilde SIA, Latvia

Spring
2003

elsnet
.....

In Memoriam Corporate Research

Annie Zaenen, Xerox PARC

In this issue, we welcome our new Opinion Column writer, Annie Zaenen from Xerox PARC in the US. Annie will be writing with us for the next year.

In ten years the research landscape in the US has undergone a major structural change: the role of corporate research has been drastically curtailed. Corporate Research as it existed in the seventies and the eighties in the States was an organisational phenomenon that is virtually unknown in Europe: research that is privately funded but open and fundamental.

Its demise is not only relevant to NLP but I'll cite only examples of research institutions which had and sometimes still have activities in that area: AT&T saw massive lay-offs a year ago, Lucent's Bell Labs have been reduced, PARC got 'spun out' and lost more than 20% of its research staff over the last one and a half years, earlier still research at IBM became much narrower in scope and Apple's attempts to create a corporate research facility doing wide-ranging research were not pursued. The only place where diversified corporate research is flourishing is at Microsoft and this points to the economic correlate that is perceived to be a requirement for the kind of corporate research known before: a *de facto* monopoly for the sponsoring companies. Even so, the research results of Microsoft are much less openly communicated than those of, say, Xerox PARC, were in the eighties and nineties. There are of course also new research efforts, e.g., at Google, but they are even farther from the open-ended, long-term corporate research of the past.

The erosion started in the early nineties and the downturn of the economy in the last couple of years has accelerated the pace. When I moved to Europe in 1993 I was under the illusion that we would be creating a research facility much like PARC. This illusion was short-lived. It became immediately clear that research had to be more closely linked to possible applications and, while I thought I would spend half of my time on benign management tasks and the other half on research, I spent most of my time fighting with concepts such as business plans, performance measures, and giving presentations to people who were completely ignorant of any technology in the field, etc. These activities have their own interesting sides, and I am not arguing that they should not be performed, but an increase in the amount of time spent on them means that, even without lay-offs, there is a reduction of research capacity. I calculate that this time increased by 15% to 20% during my time at XRCE, starting to eat

into the time not only of managers but also of rank and file researchers.

What was the reason for this shrinking role of research during a period of economic upswing? We know now that the ideology that reigned in the nineties of a return on investment of at least 15% led to all kind of contortions by old and new corporations to fulfil the expectations of the volatile investors. Among the legal ones were the closing of research labs and the attempts to make them more responsive to profit imperatives. While research labs have in general been recognized as leading to profits, this type of profit is extremely unpredictable whereas the costs are extremely predictable. So, while during the nineties the budget for R&D didn't go down, in most corporations, it was mainly the part for D that went up whereas the part for R actually shrank. Last year, however, the absolute numbers for R&D at the hundred most important technology firms fell by 6.8% according to CBMarketWatch.

In the domain of NLP proper, there was also the emergence of the Web which led to a need for retrieval tools working over enormous document collections. This created a new subfield of NLP out of IR, which up to then was considered a rather marginal part of the field. That kind of research set the standards for the whole field. Moreover the business climate created the illusion that there was high value in patents of all kinds, so intellectual property needed to be protected, another drain on research time and a barrier to the diffusion of results.

From a European perspective, the demise or drastic reduction of a handful of corporate research labs might look like a little blip on the research radar screen. Within the consistent capitalist scheme of funding research that is applied in the US, it represents an important shift. Contrary to what is the case in Europe, the US government only sponsors research in a couple of areas, mainly military and health. There is a small budget for fundamental research at the NSF. But anything that might lead to an industrial advantage should in principle be sponsored by the private sector. It is mainly to corporate research that we owe developments like Unix and the whole personal computer revolution.

Of course the ideology of not funding industrial research is in part a fiction and, especially through DARPA, the US has sponsored quite a bit of research, e.g., in NLP. But the scope of that research is controlled

e



by the sponsors. The control is more or less tight depending on the moment. In the prevailing climate, spending on NLP might well go up but it will also be tied much more closely to the needs of the intelligence community and fewer research results might become generally available. What is left of corporate research will most likely also become more secretive. When the research is done for a corporation, there are the intellectual property concerns and when research is done under DARPA or ARDA contracts, there will be national security concerns.

Relative secrecy and the narrowing focus of research are likely to slow down innovation in the US. But what is becoming a problem in the US might be an opportunity for the rest of the world. While R&D seems to be something that according to American corporations is best outsourced, new labs are created in China and in India. Nothing prevents Europe from developing a dif-

ferent strategy from the US. European firms and European public institutions should think of taking advantage of the slow down in the US to extend their sponsorship of innovative research. Europe might be in a position to replace the US as a leader in innovation. But it might also be overtaken by countries like India and China that have well-trained technicians and are starting to be in the position of making the required investments.

FOR INFORMATION

Annie Zaenen is Principal Research Scientist at Xerox PARC, USA

Email: zaenen@parc.com

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

NEMLAR: reaching out to the Mediterranean countries

New project
announcement

NEMLAR is a network project designed to consolidate knowledge about the state-of-the-art of Arabic and regional language resources, establish priority needs for industrial language technology organisations seeking to integrate Arabic and other languages into global networks, and support the development of basic resources for the partner countries and language forms.

The NEMLAR network covers recognised European centres and recognised partners in six of the Mediterranean countries covered by INCO-MED, namely Jordan, Morocco, Egypt, Lebanon, Tunisia, West Bank and Gaza Strip.

The Work Programme covers the following key tasks:

- produce a comprehensive survey of organisations, people, projects, and existing language resources for the project languages (forms of Arabic and other local languages where appropriate) and make the resulting information available as a web-based database;
- produce, by consulting Europe and Mediterranean industry representatives in speech and text technologies, a survey of observed needs for language resources for the effective development of Arabic and local language systems, and establish on the basis of this survey a set of priorities for the devel-

opment of Arabic and local language resources and tools;

- establish a LR development programme for the region, based on the observed disparity between existing resources and required priority resources, to develop a basic language resource kit covering all forms of Arabic and local languages in the region, and set realisable targets for the completion of these local tasks by NEMLAR partners, accompanied with training sessions to upgrade local personnel to LR management-readiness.

- disseminate the results of surveys, analyses, evaluations, and language resource development tasks to the human language technology community as a whole via the NEMLAR website. Hold an international conference on requirements and prospects for Arabic and other Mediterranean language resources.

NEMLAR is a joint initiative by CST (Copenhagen, coordinator), ELDA (Paris, technical coordinator), and ELSNET (Utrecht). The website will be launched soon.

FOR INFORMATION

Email: nemlar@cst.dk

Web: www.elsnet.org/nemlar

Spring
2003

elsnet
.....

Seville hosts multilingual workshop

Julio Gonzalo and Anselmo Peñas
UNED, Spain



Julio Gonzalo

A workshop on Multilingual Information Access and Natural Language Processing was held last November in Seville in conjunction with IBERAMIA 2002 (the VIII Iberoamerican Conference on Artificial Intelligence), with the collaboration of ELSNET, Ritos 2 (Red Iberoamericana

para las Tecnologías del Software en la década del 2000), SEPLN (Sociedad Española para el Procesamiento de Lenguaje Natural), AEPIA (Asociación Española para la Inteligencia Artificial), and the Instituto Cervantes.

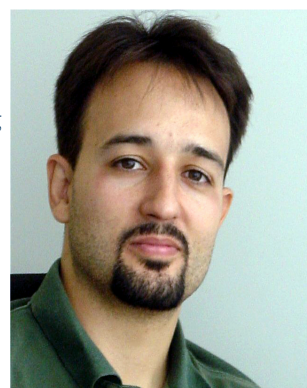
This workshop was planned as a meeting point for Ibero-American research groups in Natural Language Processing (NLP), Information Retrieval (IR), and Digital Libraries around the possibility that Language Engineering techniques and resources may help bridge the gap between the classic 'Document Retrieval' model and the broader 'Multilingual Information Access' paradigm.

The advent of the Internet and the so-called *Information Society* has quickly driven the classic retrieval paradigm into obsolescence. The term 'Multilingual Information Access' refers to the broader – and currently pretty realistic – challenge of helping users to browse, search, retrieve, recognize, and ultimately use information (rather than documents) from distributed, dynamic, and heterogeneous sources of multimedia and multilingual hyperlinked information objects. Relevant research topics include Multilingual IR, Multimedia (video, speech, image) retrieval, Interactive Retrieval, Question & Answer systems, Digital Libraries, Internet crawlers, and search engines, etc..

In this workshop, we were particularly interested in the multilingual and NLP aspects of this complex challenge, summarised in two essential issues: 'How can Language Engineering enable more sophisticated

ways of accessing information?' and 'How can we leverage Language Engineering resources and techniques to cope with large, realistic and multilingual text collections?'

The twelve papers presented at the workshop are available at the web site below.



Anselmo Peñas

These papers fall into three broad categories. The first is devoted to studies on corpora (extraction, annotation, and evaluation), which are mainly focussed on the web as a source of textual information. The second addresses a variety of natural language understanding topics, including morphosyntactic and semantic parsing and representation issues in Portuguese and Spanish, as well as extraction of lexical information (such as hierarchical relations or semantic similarity) from corpora. The final section focuses explicitly on Multilingual Information Access applications: summarisation of UNL texts, multilingual information retrieval, and multilingual information filtering based on user profiling. Most of the work presented focusses on Iberoamerican languages, although one paper addresses English, Russian, Arabic and Persian.

FOR INFORMATION

Julio Gonzalo is Lecturer and **Anselmo Peñas** is Assistant Lecturer at the Universidad Nacional de Educación a Distancia

Email: julio@lsi.uned.es, anselmo@lsi.uned.es
Web: nlp.uned.es

Web address for the workshop:
nlp.uned.es/ia-mlia/iberamia2002/program.html

e

New ELSNET members

New members

ELSNET has recently welcomed several new members into its fold. Here we give brief descriptions of four of the new members – three industrial members and one academic.

Human Language and Speech Technologies Laboratory, Sabanci University, Istanbul, Turkey

The Human Language and Speech Technologies Laboratory is one of the new laboratories founded in the Faculty of Engineering and Natural Sciences at Sabanci University in Istanbul, Turkey. Sabanci University is a recently founded private research university with an excellent research and educational infrastructure, located on the outskirts of Istanbul, Turkey's largest metropolitan area.

The Human Language and Speech Technologies Laboratory at Sabanci University was founded to pursue research and development in language engineering and speech applications, especially on Turkish. The laboratory is directed by Prof. Kemal Oflazer. Recently, Prof. Hakan Erdogan from IBM Watson Research Center Human Language Technologies Group has joined the University and the Laboratory. There are also two engineers and five graduate students working on language and speech related projects and theses.



The HLST Lab (above) and Sabanci University (below)

FOR INFORMATION

Director of the Laboratory, **Prof Kemal Oflazer:**

Email: oflazer@sabanciuniv.edu

Web: people.sabanciuniv.edu/~oflazer/

Prof Hakan Erdogan:

Email: erdogan@sabanciuniv.edu

Web: people.sabanciuniv.edu/~haerdogan/



Fourth Person Ltd, Edinburgh, UK

Fourth Person was established in mid-2001 as a spin out from the University of Edinburgh's Language Technology Group. The company has achieved several breakthroughs in natural language generation (NLG) technology, and is now developing highly advanced and innovative NLG systems for commercial deployment.

The company's language engine interfaces with any standard relational database (or other form of structured storage) to turn data into readable, speakable texts. Inherent scalability makes it far more powerful than the simple template-based language systems currently in commercial use.

Founded by Jo Calder, the chief architect of the language engine, Fourth Person currently employs three

other people, and is backed by the Edinburgh Technology Fund. The company still retains close links with the University of Edinburgh, and is located within Edinburgh's Royal Mile.



FOR INFORMATION

Email: enquiries@fourthperson.com

Web: www.fourthperson.co.uk

Tel: +44 131 525 6861

Fax: +44 131 525 6868

Spring
2003

elsnet
.....

New members
(contd.)

textkernel

Textkernel BV, Amsterdam, The Netherlands

Textkernel is developing new technology to produce advanced information extraction and text understanding engines, and to provide intelligent content management solutions for organisations with large volumes of textual data.

Textkernel's main product is Texttractor Enterprise, an adaptive multi-lingual information extraction server for the Human Resources market. Texttractor captures key information from resumes or job advertisements, and offers easy-to-integrate matchable XML-structured output.

Textkernel's products are based on a statistical language understanding and machine learning approach. This allows adaptation and development of fully operational IE systems with minimal knowledge engineering efforts.

Textkernel is a commercial spin-off from research at Tilburg, Antwerp, and Amsterdam universities and is located in Amsterdam.

FOR INFORMATION

Email: info@textkernel.com

Web: www.textkernel.com

Tel: +31 20 4942 496

Fax: +31 20 494 2499

dhaxley TRANSLATIONS, Herentals, Belgium

dhaxley TRANSLATIONS is a translation and revision company, that supplies only fully revised translations. It became a limited company under Belgian law in 1989 and gained ISO-9001 registration in 1999.

The company's mission is to supply translations of flawless quality for customers who produce a continuous flow of multilingual communications.

The permanent core team comprises language experts: translators/interpreters and philologists.

In addition to the primary specialisation in communications for the automotive and health sectors, dhaxley TRANSLATIONS provides translation of legal, financial, technical, economic, and general texts for over 200 customers, in more than twelve languages. dhaxley TRANSLATIONS maintains close links with Flemish and international centres for translation studies and translation science. By means of sponsorship, discussions and supervision of trainees, we endeavour to provide services in connection with the training of



young professional translators and revisers.

dhaxley TRANSLATIONS uses translation memory tools to increase consistency and reduce translation cost.

A consistent and accurate translation work flow is ensured by a proprietary central customer relations and job management system.

FOR INFORMATION

Email: info@textkernel.com

Web: www.textkernel.com

Tel: +31 20 4942 496

Fax: +31 20 494 2499

e

The next issue of ELSNews will include descriptions of the remaining new ELSNET members. If you are an organisation involved in language technology you may be eligible for ELSNET membership. Please check the web site www.elsnet.org for further details

Dialogue Corpora from DARPA

SIGDial

John Aberdeen, The MITRE Corporation, USA

Recently, the Linguistic Data Consortium has made available two large corpora of human-computer dialogues in the travel domain collected by the DARPA Communicator programme (see *ELSNets 104* for an overview of DARPA Communicator). These corpora are from two different data collections, one conducted during the summer of 2000 involving nine systems, and a six-month data collection conducted in 2001 involving eight systems, for a total corpus of 1,904 dialogues. The Communicator dialogues also contain user satisfaction data, based on surveys filled out by the users. These corpora are a valuable resource for the dialogue community, and present many opportunities for analysis.

In order to better understand these dialogues, various research groups have annotated subsets of the Communicator corpora, to facilitate particular analyses. For example, researchers at AT&T Labs developed a fully automatic pattern matcher to apply dialogue act tags to the system side of Communicator dialogues¹. They then built a model of user satisfaction based on several metrics including the dialogue act tags. Dialogue acts such as confirmations of what the user said were significant predictors of user satisfaction. Acknowledgements (that the system was going to perform an action that the user had requested) were also significant predictors of user satisfaction. They also found that system apologies were significant predictors of low user satisfaction.

Researchers at the MITRE Corporation were interested in comparing the Communicator dialogues with human-human dialogues in the same domain². They manually applied a variety of annotations to a subset of the Communicator corpora, as well as to transcripts of human-human (traveller-travel agent) conversations from the SRI ATIS collection and from the CMU travel collection. For example, they applied initiative annotations to a subset of the dialogues, and found that in the human-computer

dialogues the systems dominated the conversations, holding the initiative in an average of 90% or more of the turns. By contrast, in the human-human dialogues initiative was shared nearly equally by the travellers and travel agents. A related finding was evident even without the aid of annotation, namely a great difference in the number of words spoken by each dialogue participant. In the Communicator dialogues the systems uttered from four to ten times as many words as the users did, while in the human-human dialogues the travel agents uttered only about one and a half times the number of words uttered by the users.



The MITRE researchers also applied dialogue act annotations to a subset of the Communicator corpora. The dialogue act tags are nearly identical to the tags automatically applied to the system side of Communicator dialogues by the AT&T researchers, but in this case were applied manually to both sides of the dialogues, as well as to the transcripts of human-human conversations already mentioned. They found several differences in dialogue act patterns in the human-human and human-computer dialogues. For example, back-channel responses ("right", "uh-huh", "okay") were very common in the human-human dialogues, yet very rare in the Communicator dialogues. By contrast, apologies were much more common in the Communicator dialogues, which is an indirect indicator of the greater number of misunderstandings in human-computer communication.

The annotations mentioned in this article will be made available in an updated release of Communicator data from the Linguistic Data Consortium, due later this year.

References

- ¹ M. Walker and R. Passonneau. 2001. Dialogue act tags as qualitative dialogue metrics for spoken dialogue systems. In *Proceedings of the First International Conference on Human Language Technology Research*, San Diego, USA
- ² C. Doran, J. Aberdeen, L. Damianos, and L. Hirschman. 2001. Comparing several aspects of human-computer and human-human dialogues. In *Proceedings of the Second SIGDial Workshop on Discourse and Dialogue*, Aalborg, Denmark

FOR INFORMATION

John Aberdeen is Research Scientist at the Mitre Corporation, USA, primarily responsible for the DARPA dialogue project.

Email: aberdeen@mitre.org



John Aberdeen

**Spring
2003**

elsnet
.....

Language in the Knowledge-based Society: an outlook from Romania

Dan Tufis, RACAI, Romania

The 'Information Society Forum' is one of the most active professional associations in Romania, with members that are among the most representative decision makers and most highly skilled in the Romanian IT business and academic community. This forum initiated a proposal for a research study on the foreseeable societal impact in Romania of the Information Society/Knowledge Society, in close cooperation with the Section of Science and Technology of Information of the Romanian Academy. The proposal was successful, and as a consequence was included as one of the highest priority programs of the Romanian Academy for year 2001.

The program, led by the Research Institute for Artificial Intelligence (RACAI) of the Romanian Academy, engaged several top decision makers and leading scientists, and triggered a nation-wide investigation of the Romanian state of affairs with respect to the advent of the Knowledge Society. Preliminary results obtained after six months of activity were submitted to the Ministry of Research, which awarded RACAI a grant to help them finalise and publish the main conclusions of this study. As a result, RACAI published, at the beginning of 2002, a volume containing 32 studies (contributed by 44 authors) on priority themes, scenarios and possible strategies for Romanian society. One of these studies, *Promoting Romanian Language in the Knowledge-based Information Society*, which I authored, addresses the cultural dimension of the knowledge-based society. It basically argues that there is a need to complement traditional approaches to language and cultural aspects with those supported by the IT (NLP) technologies. The study urges not only appropriate joint research programmes and technical measures, but also administrative and legislative actions to foster cooperation between the scientific community on the one side, and content creators and resource holders on the other side. Nothing new essentially, but it had an impact: the first case study selected as a follow-up to this volume was Romanian language technology, present and future prospects.

The call for contribution for this follow-up volume was addressed to all Romanian language and speech scientists, and resulted in an overwhelming number of submissions. We compiled a second volume – *Romanian Language in the Knowledge Society* – with 23 papers authored by 40 specialists. The volume represents a fair overview (but inherently incomplete) of the current state-of-the-art in language and speech research for the Romanian language. Compared with a similar account on the status of language technology in Romania, published in 1997,

this volume reveals a closer connection to the international trends, good practices, and results – easily explained by an increased participation of Romanian scientists in international cooperation (research projects, thematic networks, summer schools), scientific visits, and exchanges.

With Computational Linguistics Masters programs at the

University of Bucharest, 'A.I. Cuza' University of Iasi, 'Politehnica' University of Bucharest, as well as several postgraduate courses in language and speech included in the curricula of most Romanian technical universities, the Romanian community of language and speech processing gets larger and stronger. There is of course much more to be done, but the prospects are definitely optimistic.

And Romania is definitely not a unique case. Similar activities take place in all Central and Eastern European countries, with motivations that are very similar. For several years, European networks in language and speech such as ELSNET, ELSNET-Goes-East, and TELRI (to name just a few), managed to bring together scholars with different backgrounds from all CEE countries, and paved the way towards dissemination of knowledge, techniques, tools, standards, and common practices. Moreover, various research projects explicitly aimed at languages from CEE countries (usually relying on English or French as hub languages) proved tremendously beneficial for these less spoken languages. International projects such as Multext-East, TELRI-II, or CONCEDE, usually had follow-ups, some of them supported by funding agencies in the respective countries, while others were supported in the context of new international projects. For instance, in most CEE countries, a natural follow-up development of the aforementioned projects consisted in the augmentation of language resource prototypes, to the level of wide coverage resources. To my best knowledge, this was the case for all the partners involved in these international projects. Nowadays, larger and larger corpora are available for most CEE languages, annotated with increasingly higher precision and degrees of sophistication. Wide coverage machine-readable dictionaries are also becoming available. In addition to directly serving the languages for which they were built, some of these resources became a test bed for other related multilingual studies, and often



Dan Tufis

e

constitute starting points for different types of applications. Let me give three examples, related to the Romanian language (and therefore more familiar to me).

The seven-language parallel corpus based on Orwell's famous novel *1984* was constructed by the MULTTEXT-EAST consortium. It was manually validated for various levels of annotation (sentence alignment, part of speech (POS) tagging, lemmatisation), and it represents one of the few available resources of its kind. The corpus, which was further corrected by members of the CONCEDE consortium, was largely used for various cross-lingual analyses and comparisons of POS tagging, and lately for translation equivalence extraction and lexical-token alignment. It has been downloaded by more than one hundred researchers in Europe, USA, Canada, and Asia.

A second example is concerned with the results of the CONCEDE European project, which developed a schema for multilingual dictionary encoding, and also produced a set of five core lexical databases (around 5000 entries each) encoded using this schema. These core lexical databases were/are independently scaled up by each partner, with their own financial resources. Having large Bulgarian and Romanian electronic dictionaries available in the right format, it was possible to start the DICO-East trilateral project (Switzerland, Bulgaria, and Romania), funded by the Swiss National Science Foundation under the SCOPES program. This project, based on ISSCO's DICO-PRO dictionary server, is constructing a multilingual dictionary consultation system for research and education.

Finally, the third example is the BALKANET IST project, aimed at building prototype inter-lingually aligned wordnets for the Balkan languages (Bulgarian, Greek, Romanian, Serbian, Turkish), fully compatible with EuroWordNet. Among the partners of the BALKANET project, there are two former partners of EuroWordNet (the Czech and French representatives), and the consultant of the consortium is the coordinator of the EuroWordNet project. The presence of these two former partners shows the strong determination towards compatibility with EuroWordNet (thus extending to fifteen the number of European languages represented in this multilingual semantic network). There is no doubt that the prototypes (8000 synsets per language – as specified by contractual regulations), will be further scaled up independently by each partner. The most important outcome of this project is a roadmap towards a wide coverage wordnet for each language, which has to be completely defined by the end of the project. Our own work in the BALKANET project would not be possible without the building blocks created in previous projects – the same being true for the other project partners.

At a recent meeting, aimed at preparing a project proposal for FP6, a former partner and old friend mentioned to me that after all the European projects he was

involved with he is convinced that it is much easier to work with CEE partners, compared to other partners from the rest of the world, mainly when it comes to working hard, and producing high quality results against the clock. I smiled, and I am not sure if he realised that my response was not merely complaisance. Because his words struck me as true. And there are many reasons why this is so. There is no space to further elaborate, but hopefully FP6 will offer opportunities to see why one should trust my friend's opinion.

References

Florin Gh. Filip (ed.) (2001) "*Societate a Informațională-Societatea Cunoașterii: Concepte, Soluții Și Strategii în România*", Academia Română, December 2001

Dan Tufis & Florin Gh. Filip (eds.) (2002) "*Limba Română în Societate a Informațională-Societatea Cunoașterii*", Academia Română, December 2002

Dan Tufis & Poul Andersen (eds.) (1997) "*Recent Advances in Romanian Language Technology*", Academia Română, March 1997 (available online at: www.racai.ro/books/awde)

Dimitrova, L., T. Erjavec, N. Ide, H. Kaalep, V. Petkevici and D. Tufis (1998) "Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and East European Languages" in *Proceedings of the 36th Annual Meeting of the ACL and 17th COLING International Conference*, Montreal, Canada, 315-319

Erjavec, T., Lawson A., Romary L. (1998). "East Meets West: A Compendium of Multilingual Resources." *TELRI-MUL-TEXT EAST CD-ROM*

Hajic, J. (2000) "Morphological Tagging: Data vs. Dictionaries". In *ANLP/NACL 2000*, Seattle

Erjavec, T., Tufis, D., Váradi T. (1999) "Developing TEI-Conformant Lexical Databases for CEE Languages". In *Proceedings of the 4th International Workshop on Computational Lexicography COMPLEX*, Pécs, Hungary

Popescu-Belis, A., Armstrong, S., and Robert, G. (2002) "Secure Electronic Dictionary Distribution Through the DicoPro Server: the DicoPro, DicoEast and RERO Projects". *Proceedings of LREC2002 (Third International Conference on Language Resources and Evaluation)*, Las Palmas Canary Islands, Spain, p.1144-1149

Stamou S., Oflazer K., Pala K., Christoudoulakis D., Cristea D., Tufis, D., Koeva S., Totkov G., Dutoit D., Grigoriadou M. (2002) "BALKANET A Multilingual Semantic Network for the Balkan Languages", in *Proceedings of the International WordNet Conference*, Mysore, India, 21-25 January 2002

FOR INFORMATION

Dan Tufis is Director of the Research Institute for Artificial Intelligence of the Romanian Academy; Head of the Centre for Computational Linguistics of the University of Bucharest; Professor of Computational Linguistics ('A.I. Cuza' University of Iasi and University of Bucharest); and President of the Romanian Academy Commission on Language Technology for the Romanian Language.

Email: tufis@racai.ro

Web: www.racai.ro/~tufis

Spring
2003

elsnet
.....

ELSNET Summer School – registration now open

The eleventh ELSNET Summer School on Language and Speech Communication will be held in Lille, France from 7th to 18th July 2003 and is organised by the University of Lille 3. The topic for this year's event is "Language and Speech Technology in Language Learning".

The main aims of the workshop are:

- to familiarise the students with the main principles and problems of language learning/teaching;
- to make them familiar with current best practice in computer assisted language learning;
- to make them familiar with the main challenges in computer assisted language learning

The Summer School includes courses on the following:

1. Introduction to CALL (Colpaert)
2. Introduction to NLP for CALL (Menzel)
3. Introduction to speech processing for CALL (Delcloque)
4. Using NLP tools in CALL (Menzel)
5. Using speech tools in CALL (McTear)
6. CALL design and interfaces (Delcloque)

7. Grammatical and morphological error diagnosis (Borin)
8. Systems for pronunciation training (Oster)
9. Building a grammatical CALL system (Borin)
10. Building a speech CALL system (McTear)

All students follow the same programme except for the "hands-on" sessions. Of these, students may choose one of 4 and 5, and one of 9 and 10, either NLP or speech options.

Registration for the Summer School is now open, at the web site below. The number of spaces is limited, so early registration is recommended. A limited number of grants to assist with the costs of accommodation may also be available.

Participants are expected to have a general computational background and some familiarity with language or speech research and/or processing. The working language will be English.

FOR INFORMATION

Email: elsnetadmin@univ-lille3.fr

Web: www.univ-lille3.fr/ESS2003

Web: www.elsnet.org/ess2003

ELSNET's 1000th expert

The ELSNET expert register welcomed its 1000th new member this month when Amy Isard of the University of Edinburgh signed up. Amy, who has a background in languages and a passion for the bassoon, is a Research Associate in the Human Communication Research Centre and the Institute for Communicating and Collaborative Systems in the School of Informatics at the University of Edinburgh.



She is currently working on the CRAG project, modelling alignment in dialogue, and previously participated in the M-PIRO project, which generated tailored descriptions of museum

exhibits. Amy was also a tutor at the ELSNET Summer School in Prague in 2001.

Amy's reward for being the 1000th expert on ELSNET's books is a set of three ELSNET Summer School publications.

Have you signed up yet? If you are working in any area of language or speech, join our experts' register!

FOR INFORMATION

The experts' register:

Web: www.elsnet.org/expertframes.html

Amy Isard:

Email: amyi@inf.ed.ac.uk

Web: www.ltg.ed.ac.uk/~amyi

e

Calendar

Future Events

- May 15-17** *Seventh International Workshop of the European Association for Machine Translation (EAMT) and the Fourth Controlled Language Applications Workshop (CLAW):* Dublin, Ireland
Email: workshop@eamt.org URL: www.eamt.org/eamt-claw03
- May 15-17** *Second International Workshop on Generative Approaches to the Lexicon:* Geneva, Switzerland
Email: Peirrette.Bouillon@issco.unige.ch URL: issco-www.unige.ch/gl2003
- May 28-Jun 1** *Human Language Technology Conference/North American Association for Computational Linguistics:* Edmonton, Canada
Email: lindek@cs.ualberta.ca URL: www.sims.berkeley.edu/research/conferences/hlt-naacl03
- May 30-31** *Nordic Conference on Computational Linguistics:* Reykjavik, Iceland
Email: nodalida03@hugvis.is URL: www.hugvis.is/nodalida03
- May 31** *Seventh Conference on Natural Language Learning (CoNLL03):* Edmonton, Canada
Email: daelem@uia.ua.ac.be URL: cnts.uia.ac.be/conll2003
- Jun 11-14** *Traitement Automatique des Langues Naturelles (TALN):* Batz-sur-Mer, France
Email: taln2003@irin.univ-nantes.fr URL: www.sciences.univ-nantes.fr
- Jun 16-18** *First International Conference on Meaning-Text Theory:* Paris, France
Email: mtt2003@linguist.jussieu.fr URL: mtt2003.linguist.jussieu.fr
- Jun 23-25** *International Symposium on Reference Resolution:* Venice, Italy
Email: delmont@helios.unive.it URL: www.cs.utdallas.edu/~sanda/Venice/venice-symp.html
- Jul 7-12** *41st Annual meeting of the Association for Computational Linguistics:* Sapporo, Japan
Email: acl03@media.eng.hokudai.ac.jp URL: www.ec-inc.co.jp/ACL2003
- Jul 5-6** *Fourth SIGdial Workshop on Discourse and Dialogue:* Sapporo, Japan
Email: sigdial2003@cs.cmu.edu URL: www.speech.cs.cmu.edu/sigdial2003
- Jul 7-18** *Summer School on Robust Methods in Automatic Speech Recognition:* Magdeburg, Germany
Email: wendmu@iesk.et.uni-magdeburg.de URL: iesk.et.uni-magdeburg.de/ko/summer-school/summerschool03.html
- Jul 7-18** *11th ELSNET Summer School on Language and Speech Communication:* Lille, France
Email: elsnetadmin@univ-lille3.fr URL: www.elsnet.org/ess2003

Submission deadlines

- May 1** *DiaBruck 2003:* Saarbrücken, Germany, Sept 4-6 URL: www.coli.uni-sb.de/diabruck
- May 7** *Speech-DAGM 2003:* Magdeburg, Germany, Sept 8-9, URL: speech-dagm03.uni-magdeburg.de
- May 11** *MT Summit IX:* New Orleans, USA, Sept 23-28, URL: www.mtsummit.org
- May 20** *RANLP 2003:* Borovets, Bulgaria, Sept 10-12, URL: www.lml.bas.bg/ranlp2003

This is only a selection – see www.elsnet.org/cgi-bin/elsnet/events.pl for details of more events and [deadlines.pl](http://www.elsnet.org/cgi-bin/elsnet/deadlines.pl) for more deadlines.

Spring
2003

elsnet
.....

ELSNET

Office

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

Task Groups

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

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

Research

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

Executive Board

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

The ELSNET Members: Academic sites

A: University of Vienna
Austrian Research Institute for Artificial Intelligence
(OFAI)
Vienna University of Technology
AU: Macquarie University
B: Katholieke Universiteit Leuven
B: University of Antwerp - UIA
BG: Bulg. Acad. Sci. - Institute of Mathematics and
Informatics
BY: Beloussian Academy of Sciences
CH: SUPSI University of Applied Sciences
University of Geneva
CZ: Charles University
D: Institut für Angewandte Informationsforschung
Universität Erlangen-Nürnberg - FORWISS
Universität des Saarlandes
German Research Center for Artificial Intelligence
(DFKI)
Christian-Albrechts University, Kiel
Universität Hamburg
Universität Stuttgart-IMS
Ruhr-Universität Bochum
Universität des Saarlandes CS-AI
DK: Centre for Sprogteknologi
University of Southern Denmark
Aalborg University
E: Polytechnic University of Catalonia
Universidad Politécnica de Valencia
University of Granada
Universidad Politécnica de Madrid
Universidad Nacional de Educación a Distancia
(UNED)
Universitat Autònoma de Barcelona
EL: Institute for Language & Speech Processing (ILSP)
National Centre for Scientific Research (NCSR)
"Demokritos"
University of Patras
F: Université Paul Sabatier (Toulouse III)
LORIA
IRISA/ENSSAT
Université de Provence
LIMSI-CNRS
Inst. National Polytechnique de Grenoble
GE: Tbilisi State University, Centre on Language, Logic and
Speech
HU: Lóránd Eötvös University
Technical University of Budapest
I: IRST

Consorzio Pisa Ricerche
I: Fondazione Ugo Bordoni
Consiglio Nazionale delle Ricerche - CNR
Università degli Studi di Pisa
IRL: Trinity College, University of Dublin
University College Dublin
LT: Inst. of Mathematics & Informatics
NL: Leiden University
Foundation for Speech Technology
University of Amsterdam (UvA)
University of Twente
Tilburg University
Technische Universiteit Eindhoven (TUE)
University of Amsterdam (UvA)
Netherlands Organization for Applied Scientific
Research TNO
Utrecht University
University of Nijmegen (KUN)
University of Groningen
NO: Norwegian University of Science and Technology
University of Bergen
P: University of Lisbon
New University of Lisbon
INESC-ID Lisboa
PL: Polish Academy of Sciences
RO: Romanian Academy
RU: Russian Academy of Sciences, Moscow
S: KTH (Royal Institute of Technology)
Linköping University
TR: Sabanci University
UA: IRTC UNE SCO/IIP
UK: Leeds University
University of Edinburgh
University of Sheffield
University of Cambridge
University of Sussex
University College London
The Queen's University of Belfast
University of Cambridge
University of Brighton
University of Sunderland
University of Ulster
University of Dundee
University of York
UMIST
University of Essex

Industrial Sites

B: dhaxley TRANSLATIONS
D: Acokada GmbH

D: Sympalg Speech Technologies AG
Novotech GmbH
Sasofit Aachen GmbH
aspect Gesellschaft für Mensch-Maschine
Kommunikation mbH
Verlag Moritz Dierkerweg GmbH
Langenscheidt KG
Grundig Professional Electronics GmbH
Philips Research Laboratories
Vareit Communications
DaimlerChrysler AG
IBM Deutschland
DK: Tele Danmark
E: Schlumberger Sema
Telefonica I & D
EL: KNOWLEDGE SA.
F: VECYS
LINGA s.a.r.l.
Xerox Research Centre Europe
TGID
Memodata
Aerospaciale
Systran SA
SCIPER
FIN: Kielikone Oy
Nokia Research Center
HU: MorphoLogic Ltd.
I: LOQUENDO
OLIVETTI RICERCA SCpA
LV: Tilde
NL: Compuleer
Sopheon NV
IPGibbnet Nederland BV
Knowledge Concepts BV
Texternel B.V.
PL: Neurosoft Sp. z o.o.
RU: Russian Company
ANALIT Ltd
S: Telia Promotor AB
Sema Infodata
UK: Logica Cambridge Ltd.
Sharp Laboratories of Europe Ltd
Reuters Ltd
Vocalis, Ltd.
Canon Research Centre Europe Ltd
20/20 Speech LTD
ALPNET UK Limited
Hewlett-Packard Laboratories
Imagination Technologies plc
Fourth Person Ltd.
BT Adastral Park

What is ELSNET?

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

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

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

ELSNET seeks to develop an environment which allows optimal exploitation of the available human and intellectual resources in order to advance the field. To this end, the Network has established an infrastructure for the sharing of knowledge, resources, problems, and solutions across the language and speech communities, and serving both 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. E-mail sent to elsnet-list@let.uu.nl is received by all member site contact persons, as well as other interested parties. This mailing list may be used to announce activities, post job openings, or discuss issues which are relevant to ELSNET. To request additions/deletions/changes of address in the mailing list, please send mail to elsnet@let.uu.nl

Subscriptions

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

FOR INFORMATION

ELSNET

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

Tel: +31 30 253 6039

Fax: +31 30 253 6000

Email: elsnet@elsnet.org

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

e