elsnews

11.4

The Newsletter of the European Network in Human Language Technologies Winter 2002/3

LangTech Forum 2002 – showcasing the industry

Andrew Joscelyne, EUROMAP

LangTech 2002: The New European Forum for Language Technology, was held from 26th to 27th September 2002 at the Hotel Schweizerhof in Berlin. It can fairly claim to be the first event of a new type in European speech and language circles, due to its strong focus on the commercial players and application domains of human language technologies as a whole. As such, it has been widely acknowledged as a success, with some 330 delegates, two thirds of them from 'industry' (70 different companies), coming from 20 countries. The event featured a lively exhibition of commercial suppliers, and a novelty 'elevator pitch' session with five-minute presentations from 23 young speech and language supplier companies, four of them from Eastern Europe, one from Israel, and one from Singapore.

Drs Thomas and Reuse chat with Profs Wahlster and Uszkoreit at the Opening Ceremony (© Wolfgang Borrs, Berlin)

The original idea for this industry event came from the EUROMAP project, an EC accompanying measure due to end in February 2003, that has been tracking and documenting language technology activities as R&D output has gradually been brought closer to product status in Europe since 1996. As the project's grand finale, Euromap therefore planned this event in collaboration with its German partner VDI-VDE, who worked closely from the beginning with the German organisation DFKI. LangTech was loosely modelled on SpeechTEK – the US flagship event for the voice/speech industry – even though the two events are obviously at very different stages in their development cycles.

LangTech was co-organised by Hans Uszkoreit, Head of Language Technology at the DFKI in Saarbrücken, as Programme Chair, and Bente Maegaard, Director of the Center for Sprogteknologi in Denmark and Euromap coordinator as Organisation Chair. The Local Organiser was Michael Huch of the VDI/VDE-Technology Centre for Information Technologies. The Exhibition was organised by Khalid Choukri, of ELRA/ELDA in Paris.

The IBB (Investitionsbank Berlin), which has its own programme of funding technology start-ups in the field, proved a willing and generous sponsor. The bank provided its premises for the cordial reception held on the first evening and in many other ways ensured that Berlin would be a successful location.

True to its mission, LangTech tried to cover as many aspects of the language technology marketplace as possible in two days. There were invited presentations by industry speakers, plus a smattering of academic researchers; two panels – one on various industrial issues and the other on venture capital funding of language technology start-ups; a professional market analyst's viewpoint; an SME 'elevator pitch' session to promote young companies, plus a number of demonstrations of what to expect in future.

The organisers applied a simple domain ontology to the event content, dividing its universe into 'voice', 'multilinguality', and 'knowledge management' tracks. The aim was to capture a very broad range of potential technologies and application domains. Obviously 'multilinguality' as a feature of HLT systems cuts right across any set of application domains, but this topical division allowed the 'multilinguality' track to focus attention on presentations of translation automation, independently of cross-lingual and multilingual features of speech recognition or intelligent searching that may have been cov-

Inside this Issue:

Winter 2002/3

elsnet

ISSN 1350-990X

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

© ELSNET 2002

FOR INFORMATION
Contributions to *ELSNews*, 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: 21st March 2003

ered in other tracks. Useful as this tripartite division proved, it remains a moot point as to how the mosaic-like nature of speech and language technology, whose nomenclature has been largely researcher-driven, should be terminologically tailored to the mindset of 'industrial' audiences.

Many of Europe's – and in certain cases the USA's – major players were ready to come and speak at this first-time event. To stir minds – and raise doubts – delegates were treated to the big picture in two keynotes – one by Bill Dolan of Microsoft's Natural Language Group, the other by Wolfgang Wahlster of the DFKI, head of the German-led Verbmobil programme. Both focused on ways in which language technology is being embedded inside larger IT systems.

Bill Dolan first demonstrated how language technology features (grammar checking algorithms, shallow parsing, named entity recognition, word-count summarisation) were gradually being introduced into Microsoft products as part of a strong, long-term commitment to improving consumer text processing products. But he emphasised that, while shallow techniques largely reflect the way applications have so far been defined, deeper linguistic representations will ultimately be needed for real language understanding.

Dolan then went on to show how Microsoft is working on data-driven translation, using domain specific bitexts to train a system to automatically translate very large quantities of technical documents (for Product Support Services) 'in a day'. He argued that, rather than developing a single customisable MT system, we might in future expect to see thousands of specialised MT engines scattered around the web, able to produce high quality output by pre-selecting the text domain and assigning the task to the right MT system.

He concluded by claiming that language technology's 'killer application' will not be one of these component technologies but an 'intelligent user interface' which would provide a 'natural' medium for all computer users by encompassing a whole range of language technology devices.

This prepared the way rather appropriately for the presentation by DFKI's Wolfgang Wahlster the following day on "Language Technologies for the Mobile Internet Era". Here the emphasis was strongly on the design and deployment of multimodal interfaces for mobile devices. Driven by the coming availability of 3G and UMTS technology, future mobile devices will merge such previously distinct interface paradigms as spoken dialogue, facial expression, gesture, and touch, along with the good old GUI.

These interfaces will be built on the use of face, action, lip, and speech recognition technologies, powered by sub-symbolic (neural networks) and symbolic fusion (graph unification, etc), with reference resolution and disambiguation all enabled by deep semantic representations. The idea is that the different modes can complement each other in the disambiguation process, to enable context-sensitive discourse interaction, personalisation, and situated understanding.

Wahlster then showed how this approach could be implemented in a 'transportable interface agent' – the SmartKom, instantiable as a PDA, portal, or public kiosk [*See ELSNews 11.3 for a feature on SmartKom*]. The idea is that SmartKom will enable situated, delegation-oriented dialogues – about choosing a seat at the opera, selecting or recording programmes on TV, or personal/vehicle navigation on or off the road. He noted that 8hertz, a German company demonstrating at LangTech, had already developed a versatile one-button device to physically embody the next generation of mobile devices. He also showed how DFKI itself was working with Deutsche Telekom to test and evaluate UMTS multimodal speech services in Germany, mainly in the realm of car entertainment and navigation.

What was particularly relevant about Wahlster's presentation was that, like the Verbmobil project he headed, DFKI's SmartKom programme looks ambitiously ahead towards a complete industrial nexus of activities, involving hardware, network communications, software engineering and standards development, as well as the more language technology-related activities of developing speech recognition software and semantic representations for situated dialogues.

Bill Dolan (© Wolfgang Borrs, Berlin)

This vision of large, long-term projects aimed at laying the foundations for possibly disruptive technology applications, in which language forms just one of many embedded components, in fact sounds extremely similar to the kind of vision that the European Commission is promoting in its Sixth Framework Programme (FP6) Integrated Project concept. The relevant aspects of this programme were presented to delegates at LangTech by Giovanni Varile (EC Information Society Language Unit). As is now well-known, multimodal interfaces will be a key focus in the Knowledge and Interface Calls expected in the IST branch of FP6.

After the awarding of the LangTech 2002 prizes, voted by the organisers, which went in order of excellence to Language and Computing (Belgium), Natural Speech Communication (Israel), and The Language Technology Centre (UK), it was announced by Joseph Mariani that next year's LangTech Forum will be held in Paris. *Vive LangTech!*

FOR INFORMATION

Andrew Joscelyne is a language technology analyst working with the EUROMAP consortium and was on the Programme Committee for LangTech2002.
Email: ajoscelyne@bootstrap.fr

For more information on LangTech2002 and LangTech2003:
Email: langtech2003@elda.fr
Web: www.lang-tech.org

Towards international standards for language resources

Key-Sun Choi, *KORTERM-KAIST* and **Laurent Romary,** *Laboratoire Loria-INRIA*

Note: Project names and standards quoted in this paper are fully referenced at the end of the text.

As can be seen in Cole (1997), years of research and development in computational linguistics and language engineering have led to many stable results, which in turn have been integrated into concrete applications and industrial software. At a very early stage in this rather short history, researchers and developers have understood the need to define common practices and formats for linguistic resources, since these lie at the core of any HLT work, either as a means to explore new phenomena, to parameterise pieces of software, or evaluate their results. Several projects have thus been launched to carry out groundwork on standardisation in the various domains of corpus management (TEI, Multext), multilevel annotation (EAGLES, ISLE, ATLAS), or generic software platforms (MATE, NITE).

However, none of these initiatives could reach the status of an internationally recognised standard, since there was no official standardising structure where language related activities could be considered. It is in this context that it was decided to establish, under the auspices of ISO, a new committee, TC37/SC4, dedicated to the provision of standards for *language resource management*. The aim of this committee is to build upon existing proposals in order to facilitate the development of widely reusable language resources and, further, to leverage the growth of language engineering activities.

Right from the beginning, several issues were considered as basic elements for TC37/SC4:

- We should be able to provide means of reusing linguistic data across archives or applications. This should be true at whatever level of linguistic description, from surface markup of primary sources to highly elaborate annotations at discourse level;
- In doing so, we should facilitate the maintenance of a coherent document life cycle at various processing stages, so that it becomes easy both to enrich existing data with new information and to build up complex software architectures, as long as each component can provide standard input and output interfaces;
- A clear complementarity with existing initiatives should be made explicit, so that, on the one hand, linguistic annotation can operate on a wide variety of low level format text, spoken material and even multimedia material.

Still, when we look at those issues more precisely, it seems vain to consider that it may be possible to provide fully defined formats that will deal with the various types of activities involved in language resource management. The scope of TC37/SC4 was thus centred on providing a boiler-plate for designing and implementing linguistic resource formats and processes, comprising both what is required for multi-layer annotation and the exchange of information between NLP modules. As a result, the main emphasis is put on data modelling rather than information structure design, for instance, directly expressed as an XML DTD or schema. One of our main priorities will thus be, on the one hand, to develop general data modelling principles for language resources, and, on the other hand, to provide

Working group	**Working items**
WG1: Basic descriptors and mechanisms for language resources	• Terminology of language resource management • Linguistic annotation framework • Meta-data for multimodal and multilingual information
WG2: Representation schemes	• Structural content (morpho-syntax and syntax) representation scheme • Multimodal meaning representation scheme • Discourse level representation scheme
WG3: Multilingual text representations	• Translation memory and alignment of parallel corpora • Segmentation and counting algorithms • Meta-markup for globalisation, internationalisation, and localisation
WG4: Lexical database	• NLP lexica representation scheme
WG5: Workflow of language resource management	• Validation of language resources • Net-based distributed cooperative work for the creation of language resources

Table 1: Overview of planned working groups and working items

means for existing projects to describe their own formats under those principles, so that automatic mappings are generated from those descriptions. Of course, there will be a need for newcomers to have baseline representations to start their work, but such reference representations will only be, for a given type of activity (e.g., dependency syntactic annotation), one member within a wider family of activities (see Ide and Romary (2001) as an illustration).

In the context presented above, a preliminary workplan has been designed, which organises TC37/SC4 along a group of five potential working groups. These working groups, as well as the working items that they could comprise, are described in Table 1 above.

There are obviously topics which are not explicitly covered in this general workplan and which deserve attention, either because they correspond to deep needs in language engineering or because there actually exists a mature proposal which could serve as a basis for future standards. There are in particular a few topics clearly missing. We can mention, for instance, all the basic statistical methods that are used either in stochastic models or for evaluation purposes.

In this respect, TC37/SC4 has to adopt an opportunistic strategy by fostering the creation of new work items (and amending the workplan accordingly) whenever there seems to be a good consensus on a given topic. Comments and suggestions are thus welcome to make this workplan more accurate and conformant to the needs of our community. From an administrative point of view, the new sub-committee Four (SC4) on language resource management has been established within an existing technical committee named TC37. This committee has been working for more than 50 years on the definition of standards in the field of terminology and has provided many reference documents in this domain. In recent years, it has gained some experience in the development of standards intended to facilitate the management of computerised terminology (e.g. ISO 12200, ISO 16642). It has also worked on standards whose aim was obviously wider than just terminology work, as can be seen with the two standards on language code (ISO 639-1 and ISO 639-2)

As an institution, the ISO only works because it represents its various member bodies, and organises the technical work for them. A member body is a national standard organisation that has full access to technical activities, will vote and/or comment on whatever documents are circulated, and above all send experts to working groups and committee meetings. As a consequence, people willing to be active within TC37/SC4 should make themselves known to their respective organisation (e.g. DIN in Germany, ANSI in the USA, BSI in UK etc.).

Beyond national participation, TC37/SC4 will not work unless there is strong and active support of the major associations and networks such as ACL (through its dedicated SIGs such as SIGDIAL, SIGSEM, SIGPARSE in particular) and ELSNET. Those structures are the only ones to provide both the right level of expertise during the design phase of standards and the proper dissemination environment that will ensure that standards are actually understood and used.

On the whole, we think this is a major opportunity to stabilize technologies and formats which people have been working on for many years so that we encourage future work and development in language engineering that go far beyond the current state of the art. At the same time the responsibility lies upon all of us in our community to contribute to this new initiative, so that future standards will both be of high quality and correspond to real industrial and research needs.

FOR INFORMATION

Key-Sun Choi is Director of KORTEM-KAIST, **Email:** kschoi@cs.kaist.ac.kr, **Web:** kschoi.kaist.ac.kr
Laurent Romary is researcher at Laboratoire Loria-INRIA, **Email:** Laurent.Romary@loria.fr, **Web:** www.loria.fr/~romary

References

Cole Ronald A. (Editor in Chief), Joseph Mariani, Hans Uszkoreit, Annie Zaenen,Victor Zue (Eds), 1997, Survey of the State of the Art in Human Language Technology, 1997, CUP (also available online at cslu.cse.ogi.edu/HLTsurvey).
Ide N. and L. Romary 2001, A Common Framework for Syntactic

Normative references

ISO 12200: Martif, Machine Readable Terminology Interchange Forma
ISO 16642: TMF, Terminological Markup Framework.
ISO 639: Code for the Representation of Names of Languages.
ISO 639-2:1998, Code for the representation of names and languages-part 2:Alpha-3 code.
XML (Extensible Markup Language) 1.0 (Second Edition) W3C Recommendation 6 October 2000, Tim Bray, Jean Paoli, C. M. Sperberg-McQueen, Eve Maler (Eds.).

Pointers to projects or initiatives

ATLAS (Architecture and Tools for Linguistic Analysis Systems), www.nist.gov/speech/atlas/
CES (Corpus Encoding Standard), www.cs.vassar.edu/CES/
EAGLES (Expert Advisory Group on Language Engineering Standards), www.ilc.pi.cnr.it/EAGLES/home.html
ELSNET (European Network of Excellence in Language Technology), www.elsnet.org/
ISLE (International Standards for Language Engineering), lingue.ilc.pi.cnr.it/EAGLES96/isle/
IMDI (EAGLES/ISLE Meta Data Initiative), www.mpi.nl/ISLE/
ISO (International Organization for Standardization), www.iso.ch
LISA (Localization Industry Standards Association), www.lisa.org/
Mate (Multilevel Annotation, Tools Engineering), mate.nis.sdu.dk/
MPEG (Moving Picture Experts Group), mpeg.telecomitalialab.com
Multext (Multilingual Text Tools and Corpora), www.lpl.univ-aix.fr/projects/multext/ (see also Multext-East, nl.ijs.si/ME/)
Nite (Natural Interactivity Tools Engineering), www.dfki.de/nite/
OAI (Open Archives Initiative), www.openarchives.org
OLAC (Open Language Archives Community), www.language-archives.org/
OLIF (Open Lexicon Interchange Format), www.olif.net/
OMG (Object Management Group), www.omg.org/
OSCAR (Open Standards for Container/Content Allowing Re-use), www.lisa.org/oscar/
SIL (Summer Institute of Linguistics), www.sil.org
TEI (Text Encoding Initiative), www.tei-c.org
W3C (World Wide Web consortium), www.w3.org

Canoo announces new German language products

Canoo Engineering AG announced the release of its Word Manager Transducer (WMTrans) product range in November 2002. The German morphological analysis software offers intelligent text processing for information retrieval and language processing applications. Typical uses include intelligent search, text indexing, text mining, language learning, hyperlink generation, spell checking, grammar checking, and machine translation.

The tools that comprise WMTrans are all based on Canoo's German Morphological Dictionary, containing more than 200,000 entries and generating over two million fully categorised word forms, including information on word formation, all types of inflectional irregularities, and spelling variants.

ELSNews spoke to Stephan Bopp, responsible for the development of much of the linguistic framework used in the dictionary. He explained, "The morphological database is rule-driven, with the lexeme as the central entity. Rules define regular inflection and derivation, while irregular and subregular forms are provided in full. Lexemes all belong to an inflectional paradigm, which defines the stem and suffix possibilities, and the grammatical features that are represented by the stem and suffix combinations. In addition, a transducer determines the relation between underlying combinations and surface forms, taking into account, for example, spelling rules." Bopp pointed out that the morphological dictionary differs from the CELEX database, which covers largely the same information, in both its structure and coverage. "CELEX is much flatter, while WMTrans has more structure and rule-based information. As a result, CELEX is much less complete, as the WMTrans dictionary makes it possible to recognise and categorise forms that have never been encountered before."

As well as a rule-based treatment of inflection, the WMTrans toolkit includes a grammar for compounding and derivation, both of which are extremely productive in German. "These rules define how compounding and derivation form new stems and which inflectional paradigm the stem belongs to", said Bopp. "It is the treatment of derivation and compounding that is really novel in WMTrans."

Sandra Wendland, head of marketing at Canoo, said "I would like to emphasise the fact that we offer WMTrans as a commercial product. One significant area of application is Information Retrieval. Intelligent search tools for the Internet and desktops are becoming increasingly important. Our reference customer list includes Google and NZZ Online.

"Applying WMTrans saves developers time and money. Using the API for Java or C/C++, the required functionality can be easily added to Java or C/C++ applications. The WMTrans products can be integrated into existing or new applications. Developers can concentrate on building the application and rely on WMTrans products to compute the relevant data."

Stephan Bopp

The range of tools available includes the following:

Lemmatiser – returns citation form and major category for a word form;
Unknown word lemmatiser – used if the lemmatiser fails to recognise a word form such as a compound, returns citation form and major category;
Inflectional analyser – determines the base form and category of a word form as well as providing additional grammatical and orthographic information;
Recogniser – detects if a character string is a valid German word (according to morphological and spelling rules);
Inflection Generator – returns all inflected word forms and spelling variants for a base form;
Inflection analyser/generator – determines the base form and category of a word and computes all possible inflected forms and spelling variants for that base form;
Word formation analyser/generator – determines the components of a derived or compound word and finds all possible word composites and derivations in which it is involved.

The two lemmatisers are both available for Windows, Linux, and Solaris. The other tools are currently available for Linux only, and free trial dowloads are available.

FOR INFORMATION

The Word Manager Transducer (WMTrans) Product range is available from Canoo Engineering AG.

Contact: Elisabeth Maier
Canoo Engineering AG
Kirschgartenstr. 7
CH-4051 Basel
Switzerland

Tel: +41 (61) 228 94 44
Email: wmtrans-info@canoo.com
Web: http://www.canoo.com

Opinion

***Kenneth Church,** AT&T Labs Research, USA*

This will be my last *ELSNews* column. It has been great, but now it is someone else's turn to have their say. I have enjoyed writing these columns, especially when I have had the time to send drafts around the world for feedback. My previous columns have benefited from a lot of people I want to thank: Amir Amihood, David Johnson, Mark Liberman, and many others. Unfortunately, I find myself writing this last column up against a deadline. I know it is my own fault, but that doesn't make it feel any better. I'm sure you can all relate to that. No time for help. No time for fun. No time, period.

I had planned for some time to write my final column on how to write a great paper without great material. This is actually a hard topic to write about, which is part of the reason why I have procrastinated for as long as I have. I see content and impact as separate dimensions. Papers can have more or less content and more or less impact. Content is an internal property of what is in the paper itself, whereas impact addresses how the paper is received by the external community. Some papers stand up well to the test of time and some don't. Many of us find it easier to focus our energy on the content dimension and hope that that effort will somehow (by magic) make progress along the impact dimension. I want this column to take a different approach and focus exclusively on the impact dimension. In particular, is it possible to write a content-free paper that makes a difference?

Part of the reason that I am up against the deadline (not that I have any good excuses) is that I will be going to a thesis defence in a few days, and I should be reading the thesis right now (another thing to feel guilty about). The thesis is actually quite good. Students usually don't get to schedule a defence unless it has already been decided that they are going to pass. So, if I can't help (much) with the content, what I am doing at the defence? Of course, it is important to celebrate the event. But in addition, I can still help with impact, even at this late date.

The thesis is on information retrieval (IR). It is ironic that IR is about how to find stuff, and my job, as a member of the committee, is to help with impact: that is, how to make it so the community can find this stuff (and remember it).

There is quite a bit of prior art in how to fool search engines. There is the standard trick, for example, of writing "sex" a million times where people will never see it (in tiny 2 point font, off the margin in a color that will blend into the background). Of course, there are a few problems with that. First, search engines have wised up to the standard tricks, so those tricks won't work any more. And secondly, I have my doubts about the effectiveness of crass marketing tricks (spam). They probably don't work in any case, but they are surely off-limits for selling a thesis.

That is not to say you shouldn't sell a thesis. Marketing is all about understanding the audience (the market). Think about what you can do for them and how you can measure their reaction, so you can see if you are making progress or not with what you are doing. Evaluation metrics are very much in vogue these days. The well-established review process and much of what I learned in graduate school is focused on content. There are also well-established metrics that focus on impact such as citation indexes. If the work gets picked up in secondary sources (textbooks and review articles), that's also an indication of impact. Readership matters. The publishers know which authors sell. Name recognition counts.

The modern web context has adopted (and improved) many of these well-established metrics. Readership can be measured in terms of page hits (and buzz). Most publications are never cited. Few are even read. Statistics on page hits will help us understand these grim realities, and might even offer some constructive suggestions of what we can do about it.

Graduate students have always been concerned that no one (not even the committee) will read the thesis. My wife just reminded me of the old trick of putting a hair in the copy that you give to your advisor to see if he actually read it. I did something similar; see if you can find the word "tunafish" in my thesis.

What can students do to increase readership (hits)? Search engines these days not only make use of internal content (words) of a document (e.g., scoring functions like td*idf), but they also make use of external properties (e.g., the Google page rank algorithm looks at URL links from the community to the document in question). Tricks like the "sex" trick mentioned above can fool content-based retrieval engines but they don't work as well on impact-based retrieval engines because you can fool some of the people some of the time, but it is harder to fool much of community much of the time.

I am very impressed with citeseer.nj.neccom, a citation index of online research papers. It may not be as appropriate as well-established citation indexes for important promotion decisions because the sample of papers is not carefully balanced, and because the sample includes a lot of papers, some of dubious quality (e.g., unpublished papers). Nevertheless, sample size can make up for a lot of problems with balance. (More data is better data. See my previous column in *ELSNews 11.3*.) Whenever I run CiteSeer on the papers in an area that I know something about, the more important

➡

papers (and authors) are almost always at the top of the list. In fact, CiteSeer does at least as good a job as textbooks in identifying the important stuff in an area. I regularly use CiteSeer in areas that I don't know to identify who's who and what I ought to read.

So what can students do to increase citations? Obviously, having great material (content) helps, but I want to focus on other things. There are a few things that help that I won't talk about (because they are hard and take time): good writing, an entertaining talk, lots of friends in the right places (the old-boy network) and saying nice things about as many people in the community as possible.

Make it easy to find your work. Apparently, papers that are available online are more heavily cited ("Online or Invisible?", Steve Lawrence (*Nature*)). I am concerned that this analysis is self-serving and biased (the citation index used in the study was based on a collection of online articles which the author of the study was involved in). Even so, the conclusion is undoubtedly correct. Distribution matters. Publish early. Publish often. Publish in the "right" places.

Some small (content-free) changes to the write-up can make a big difference to citations:

1. Publishing data: I ran CiteSeer on "corpus" and it came back with the Penn TreeBank, the Wall Street Journal Corpus and Switchboard. Despite the fact that reviewers sometimes balk at publishing data (the reviews of the Penn TreeBank were brutal), these publications have been good for the field and the field returns the favour with lots of citations. I often recommend that students publish URLs to whatever useful data they can. If anyone else gets any value out of it, the student will almost always get a citation in return. This has at least been David Lewis' experience with Reuters.
2. Publishing tools: I ran CiteSeer on "toolkit" and it came back with even more citations than "corpus"! In my own experience, if you distribute a useful tool (like a part of speech tagger or a tool for aligning parallel corpora), you will receive lots of citations whenever the tool is used. Think of citations like royalties.
3. Helping students get started in the field: tutorials, bibliographies, appendices that can be used as homework problems. Students will cite the place where they first learned about something. (Secondary sources such as textbooks and review articles are heavily cited.)

In short, I have suggested that impact (citations) is more important than content (peer review). Doing great work is great, but it is even better if it makes a difference. In many businesses, sales and marketing is considered at least as important as R&D. We ought to think of citations as a measure of sales. R&D is also important, but largely as a leading indicator of future sales.

FOR INFORMATION

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

Email: kwc@research.att.com
Web: http://www.research.att.com/~kwc

Some new opinions

Although it is sad to say goodbye to Ken, who has given us plenty of food for thought during his tenure as our Opinion column writer, we are delighted to welcome in his place Annie Zaenen, who will pick up the reins in the next issue.

Annie started her career in Belgium, taking her Kandidatuur and Licentie at the Rijksuniversiteit in Ghent. She was awarded a Ph.D. from Harvard in 1980. She has worked for the Xerox Corp. since 1985, having been Area Manager for Natural Language at XRCE-Grenoble up to 2000. She is currently principal scientist at the Palo Alto Research Center. Annie has a wide ranging experience, having also worked for the Ministry of Public Education (Belgium), Centre de Recherches et d'Information Socio-Politiques (Brussels), University of Geneva, Rijksuniversiteit Ghent, Harvard University, University of Pennsylvania, Massachusetts Institute of Technology, and Cornell University. She has been on the editorial boards of *Linguistic Inquiry*, *Language*, and *Natural Language and Linguistic Theory*, a member of the EAGLES board, and one of the editors of *The State of The Art in Human Language Technology* (NSF-CEC), and is currently on the board of ACL.

Annie says of her own research interests:

"I am a linguist and most of my own work is in theoretical syntax. However, it has always been important for me that linguistic descriptions be embedded in implementable models. The work that I have done and do with colleagues at PARC is aimed at ensuring this. While I use deep parsing models in my own linguistic work, I pushed for the development of shallow techniques as the manager of the XRCE NLG in Grenoble because these approximations give us insights into what are the most prevalent cases of linguistic dependencies. In my current research I collaborate with colleagues at PARC to promote the integration of machine learning techniques and formal grammar and with colleagues at Stanford to make corpus analysis into a necessary tool in theoretical syntax."

Annie Zaenen

From phonetic symbols to the Cambridge Grammar

Lynne Cahill** interviews **Geoffrey Pullum

GEOFFREY K. PULLUM is a linguist specialising in the study of English, and has published widely on linguistics. He holds a B.A. in Language from the University of York (1972) and a Ph.D. in General Linguistics from the University of London (1976). Between 1974 and 1981 he taught at University College London, the University of Washington, and Stanford University. He was a Fellow of the Center for Advanced Study in the Behavioral Sciences in 1990-91. Since 1981 he has worked at the University of California, Santa Cruz, where his title is Professor of Linguistics.

He has published a dozen books and nearly 200 technical articles within the field of linguistics. He was co-author of the book Generalized Phrase Structure Grammar *(1985) and coeditor of the four volumes of the* Handbook of Amazonian Languages *(1986-1998). Perhaps the best-known of his books is* The Great Eskimo Vocabulary Hoax *(1991), a highly entertaining (and often very funny) collection of satirical essays about the field of linguistics that originated as columns in the Topic...Comment series in the journal* Natural Language and Linguistic Theory. *He is currently travelling the world promoting his most recent publication, the massive* Cambridge Grammar of the English Language, *co-authored with Rodney Huddleston. ELSNews caught up with him on a visit to Sussex.*

LC: In writing the Cambridge Grammar, do you think that your respective backgrounds give you good coverage of the varieties of modern English?

GKP: Yes. We figure that Rodney was born in Manchester, grew up speaking a Northern dialect of British English, was educated in Cambridge and Edinburgh, then moved to Queensland and spent 25 years living in Australia. So he's very well acquainted with British English, both from the North and the South and Australian English. I spent nearly 30 years living in England, where I was born, and then moved to the US in 1980, permanently as it turned out, and that's 22 years of experience with American English, which has become my adopted language. I actually lecture in American or at least Americanised English and I've become much more used to its features. So we felt that gave us enough coverage. You could say that there should have been an American on the full-time main authorial team, but we were getting advice from Americans. For example, Jim McCawley was a consultant before his untimely death; Geoffrey Nunberg was a co-author of a chapter; Gregory Ward and Betty Birner were co-authors on another. They and lots of others looked at the book.

LC: Do you find yourself looking out for strange constructions?

Geoffrey Pullum

GKP: No. Wittgenstein, I think, said "Thinking isn't something you do, it's something that happens to you" and that's what it's like with linguistic investigation. I'm powerless to control this but things strike me now much more than they did and so I find wonderful stuff everywhere, but it's not because I'm looking for it. It just comes to me and squeals and howls at me to be noticed and I'm often open-jawed in amazement at things in books that I'm reading or that people have said. Discovering that "bush" was a directional preposition in Australian English was a real thrill and I still get pleasure from a well-turned split infinitive or a nice singular "they" with a definite NP antecedent, which is delicious. But you don't have to actively look out for it. I actually don't keep a notebook on me to note things down, which is a problem, but I have taken to using five or six bookmarks. The first five are for marking interesting examples I've noticed on some pages and the sixth one is for keeping my place.

LC: You have said that your ambition in writing this book was "to change the common vocabulary". What do you think are the chances of achieving that?

GKP: I don't know how to estimate the chances. What we're up against is the conservatism of two professions at least: English language teaching and dictionary making. I don't know how conservative they are. We may be up against more than that. If it is true that by law in the UK primary school children are being taught to use the Quirk terms for things then we're up against the law of the land as well as the inherent conservatism of language teachers and dictionary makers and it will be tough. But it's worth having a shot at it. One shouldn't give up on things just because they are difficult.

LC: You have said a lot about how the Cambridge Grammar improves on the Quirk et al. (1985) volume. How does it improve on the Biber et al. volume?

GKP: Biber et al.'s book, *The Longman Grammar of Spoken and Written English*, which we studied carefully, of course, is a thinner description than Quirk. It gives much less detail and on some points gives not very much detail at all. The bulk of the book and its innovation and additional information is the statistical tables it has. We don't really regard that as grammatical information. That is, it is a fact about the verb "approve" that you hear it sometimes with an "of " phrase complement and sometimes with a noun phrase complement and they have different meanings but it's not really a grammatical fact what the percentages are of the occurrences of the two in private letters to friends as opposed to published fiction. That's not to say that information is not useful – I turn to the Biber volume when I want to know what is the percentage of clauses in modern conversation that begin with "whom" – it turns out to be zero percent. It's interesting and they do have the figures. It's useful occasionally for that, but it's not a new grammar. The Cambridge Grammar is not really corpus based in the sense of fixing on a certain collection of texts and depending on everything in there so that if something did not occur there it was left out. There was no corpus in that sense, which again is not the same as saying we didn't bother to collect real evidence. We always went for real evidence unless the thing to be illustrated was absolutely trivial and "the cat sat on the mat" would do. We're interested in describing modern standard English as it is, but no corpus will suffice for that. So we used a back and forth procedure between consulting our intuitions, comparing with the intuitions of other people we regard as speakers of the same language, getting corpus evidence to illustrate what we think is the point, modifying what we think our description should say when we see what the corpus evidence reveals and rechecking against our intuition. This process is in fact reminiscent of what John Rawls, the political philosopher, describes as "the process of obtaining reflective equilibrium". That's really what a grammarian has to do and any methodology that said "just by the corpus, strictly by the corpus and nothing but the corpus" would be an oversimplification. Of course we were very suspicious of the tendency of some syntacticians to rely on intuition absolutely to excess. Neither of these two paths alone covers how we did the Cambridge Grammar and I would have thought that what we did is a solid middle ground where many people could agree.

LC: Do you think more intensive uses of corpora are appropriate for some areas of linguistics?

GKP: Especially in computational linguistics, yes, and in dictionary making. Suppose you are trying to write the different senses of a word. It's an excellent idea to write them according to the frequency in a corpus you hold fixed for the whole dictionary. My Concise Oxford English Dictionary, which I won in a competition in 1957, gives only one meaning for "awful", that is "tending to cause or inspire awe or wonderment". Now that's a bit out of date. Were you to do it today, how would you decide whether to put the "bad" sense of "awful" before the one in "his awful majesty"? Well I think the relative frequency in a modern corpus is an excellent way to choose.

LC: What about the use of the web as a corpus?

GKP: It's a great idea. It seems to work much better than I would have thought. My student, Chris Potts, is a real determined enthusiast of real data. He likes to use real examples to illustrate points, even when they are difficult and subtle and without any special search technology, just tracking down quoted phrases from Google he does great things with the web. It's much better than I would have thought it was. I would have thought there's so much junk on people's web pages, so much misspelled ungrammatical rubbish that the results would be almost unusable. But it's not, it's great. It's a really wonderful opportunity that couldn't have been dreamed of ten years ago when the Cambridge Grammar had already started. Google is the most spectacular case and what they have done is a staggering achievement of the technology. It's a huge bonus to linguists of all stripes, but I can well understand why computational linguists are now making more use of the web as a corpus than the BNC.

LC: Gerald Gazdar has talked in an interview with Ted Briscoe of how you four (Gazdar, Pullum, Klein, and Sag) believed that you could change the field when you wrote the GPSG book. He no longer believes that and has suggested that you probably don't either, even though you manage to appear as though you do. Is that a fair assessment?

GKP: I am in fact as optimistic as I appear to be. It's not a political stance. I'm very happy interacting in the American linguistic context, puzzling and frustrating though it may sometimes be, because of its great sensitivity and sociological and political factors. The simple fact is that Johnson, Lappin ,and Levine are right that the swing to minimalism is baffling on rational grounds because it cannot be motivated by anything empirical or theoretical, but the whole of mainstream American syntax has taken this wrong turn because they're following Chomsky. The explanation lies in his personal prestige. That is a salient fact about the character of the American intellectual establishment that it can be swayed off course that much, but to me it's not a case for pessimism or cynicism. It doesn't mean there is no truth, it means that, in the US, a prestigious charismatic figure can swing 90% of the public away from the truth for a while. But it doesn't last and in a country of 280 million people you can always find several brilliant people cleverer than you who are doing stuff of just the sort you think should be done and learn from them. The great thing about a country that big and that diverse is that wherever your interests lie, there is an interest group of people who are like thinkers and you won't be the smartest person in it. That's probably a paradox, but it seems that way to me. So I have a great time in American linguistics. All sorts of things about it change all the time and the thing that's going to prevent you changing it ever is to lose your faith that change is possible. When I say that I think that people can be won over and changes can be made, it's not a stance. I'm not pretending that I think that. I think that and indeed I have personally swung things round. Tiny things but there was a time when it looked as though we were in real danger of having Luigi Burzio's ridiculous terminology for verbal classes widely adopted and I smashed it and pushed people back in the direction of the sensible terminology that Perlmutter and Postal had worked out. Their terminology (ergative, unergative, accusative and unaccusative) was a good classification and was in use for a while. Then Burzio started calling unaccusative verbs ergative verbs, which is an absurd usage not supported

➡

by anything. I managed with one Topic..Comment column to swing things round. So yes, you really can change things. You don't have to follow fashion to have interesting and useful things to do in a community as big as the USA. You don't even have to do without intellectual company.

LC: Do you think your academic path would have been very different if you had not moved to the US?

GKP: Yes. The move to the US had enormous effects. The main thing really was the energy boost obtained from operating in a context that much larger and more competitive. There are areas of the world where conferences are run in little communities of academics and the rule for acceptance of papers is, if somebody offers something, they should be allowed to present it at the conference. This is not followed in the USA. You can expect 60% rejection rates for meetings. It's tough. You've got to craft your abstract well and revise it five times and still cross your fingers. Then if you get in it's actually an achievement. That to me is energising. The energy level created by this more competetive environment is very important to me and I think in all sorts of ways it has shaped what I've been able to do. That's true in general linguistics, where the major community is the Linguistic Society of America and in computational linguistics, where the ACL is continually more professionalised and demanding It's also only in America that I've encountered real transfer to industry. That's started up since I left Britain. When I left in 1980 it was just getting started in Silicon Valley and I'm very glad I was there to participate in that during the majority of the 1980s, when HP was doing research in GPSG building a NL access system for databases I'm not sure what happened in the late 1980s to convince HP to reassign everybody on that project to handwriting recognition – maybe they saw the unintelligible writing on the wall! I think we should have concentrated harder on getting the first serious commercial system up and running with emailed plain English text going in and accurate answers coming back. I still think that's something we've got to return to. It's crazy to go on with computers of the power we have now with us trying to learn their languages when they should be learning ours.

LC: Was it enjoyable working on the Cambridge Grammar as a change from more formal work?

GKP: In fact I have not left formal linguistics behind, it's a parallel strand of my work in which I'm trying to apply model theory to syntactic description. It's quite technical and radically opposed to the generative approach. It's compatible with the Cambridge Grammar because there's nothing about the Cambridge Grammar that's based on the notion of generative mechanisms. You won't find anything like movement rules, not even disguised in informal terms. We give descriptions of sentences in terms of statements in a logic that can be satisfied by some structures and not by others. The grammatical structures are the ones that satisfy the statements of the grammar and the relation between sentence structures and grammars is the satisfaction relation of model theory – not "is analagous to" or "is like" but *is* that relation. Barbara Scholz and I are teaching a foundation course on it at ESSLLI this coming Summer 2003.

LC: What would you say is your favourite achievement? One of my own personal favourites is the Phonetic Symbol Guide.

GKP: Interestingly, it was a completely satisfying thing to do and it's a book I keep close and need to refer to often. I had the strange experience while doing the work for the Phonetic Symbol Guide of perceiving a kind of inversion between my perception of kinds of research and the standard one. The standard view is that, while humanities people speculate and think about poetry and there's really no truth, scientists are steadfastly marching forward and gathering up truth. I began to see as I worked on Phonetic Symbol Guide that this was not the case. If we found an 1896 use of a symbol and previously we thought it was invented in 1903, we had a certain fact that there was an earlier source. There was a calm certainty about humanities research of that kind that scientists knew nothing about. Science is actually a process of swirling speculation and obscure conclusions, experiments that can be interpreted in more than one way. No scientist worth her salt ever really knows what's going on. It's exciting, but truth and certainty you don't get.

However, of all my career, the Cambridge Grammar is without question the thing I am most proud of being associated with. I should say that Rodney Huddleston did twice as much as I did on that, at least, but I'm proud just to have worked along with him. In the end, I'm rather seriously inclined, despite my tendency to joke. I enjoyed enormously writing the essays in the Great Eskimo Vocabulary Hoax but they're not serious work, they were never intended to be. It's depressing to see them taken too seriously. That was enormous fun, but I am by nature rather serious and I want to do things for real. The Cambridge Grammar is a real contribution on a level I had not previously reached and therefore more satisfying to be involved with than GPSG which was of course, being real science, shrouded in speculation and confusion. At the end of writing a whole book we still didn't know anything about what's really true. That's probably why there was no concerted effort to sell the ideas of GPSG in 1985 and the following years. Even the authors couldn't be sure that anything they said was true. That's the thing about the Cambridge Grammar – the things in there are true. It's a new level of job satisfaction.

FOR INFORMATION

Geoffrey K. Pullum:
Email: pullum@ling.ucsc.edu
Web: ling.ucsu.edu/~pullum/home.html

The Cambridge Grammar of the English Language:
Web: www.cambridge.org/cgel

ESSLLI Summer School 2003:
Web: www.logic.at/esslli03

Summer School report

European Masters in Language and Speech is booming

Compiled from reports by Peter Dirix and students from UPC Barcelona

60 participants in Leuven summer school

Started as an ELSNET initiative, fourteen universities co-operate in a common framework of a European Masters in Language and Speech. Every summer, lecturers and MA students from the programme come together in a summer school. This year, 60 participants from universities in the UK, Belgium, the Netherlands, the Czech Republic, Spain, Greece, Germany, Switzerland, and Poland came to Leuven for a school, organised by the Centre for Computational Linguistics.

Some student impressions

(UPC Barcelona students)

After a hard working semester, summer arrived and a good opportunity faced us. It was to attend the European Masters Summer School that this year took place in Leuven, Belgium. Even though people were tired of studying and/or working hard, this seemed a very good chance to improve our knowledge of language and speech technologies, so we decided to give it a try.

Through that week we attended the tutorials we had chosen. In a short time you can get an overview of Grammar Formalisms, Physiology, and Morphology or get introduced to Speech Synthesis, VoiceXML, and Automatic Speech Recognition. The one bad point was that students were not able to attend all the tutorials they wanted. From our point of view, practicals were useful and easy to follow even if you didn't know anything about the specific topic. The best thing about the EM School is meeting other students around Europe and also researchers from relevant research laboratories who can help you with your current project, telling you about the real projects that are currently going on and prospects for the Language Technologies.

Leuven was an appropriate town to hold the EM Summer School. It was calm and peaceful during the day, but at night we could enjoy many night-life activities. The most enjoyable activity was to sit in a bar and taste some of the delicious Belgian beers. Another remarkable thing about Leuven was its proximity to Brussels. The excursion to that city was interesting but it was a pity that it was a cloudy and rainy day (the weather is maybe one of the disadvantages of Belgium, although during the week it was not too bad). We think it is a good idea to encourage people to travel around the place where the EM School takes place.

The organisation in general was good, but it would have been better if accommodation had been provided for all the students. Summing up, our global evaluation about this Summer School is really positive. We would like to thank ELSNET for the support given and also to suggest that they keep the grants for the EM School on for future Summer Schools.

A student's diary

(Peter Dirix)

On Monday, we started with registration, followed by a short introduction and a lecture by Gerrit Bloothooft about voice source characterisation. The rest of the day was filled by students presenting their final projects. It was a nice mixture of speech and language technology. On Tuesday and Wednesday, eight tutorials took place. Most of the Flemish students took Karel Pala's tuto-

➡

rial on building corpora and constructing a morphological analyser. We did this for Dutch, French, and Spanish. There were also tutorials by Martin Cooke, Simon King, Ivan Kopecek, Martin Rajman, Michael Moortgat, Dirk Van Compernolle, Astrid van Wieringen, and Vincent Vandeghinste. Most students seemed to be very happy with the ones they took.

On Thursday 11th July, the Flemish national holiday was celebrated. After some more student presentations in the morning, an excursion to nearby Brussels was planned. The Flemish government had organised some huge festivities in that city to celebrate the 700th anniversary of the Battle of the Spurs in Kortrijk. First, the Leuven students took the foreigners for a tour through the city. We showed them the cathedral, the Grand'Place, the Parliament, the Bourse (stock market) and of course Brussels' most famous citizen, Manneken-Pis. After dinner, some street concerts and the big show on the Grand'Place were visited. It was really a pity we had to take the last train back to Leuven. Friday was the last day of the summer school. In the morning, the last tutorial sessions took place. In the afternoon, the students had the chance to present the work done in the tutorials to the whole group. Some students had already left by then. The remaining students went for dinner on Leuven's Oude Markt and then had some drinks (Belgian beer, of course!) while enjoying the jazz sessions of the Beleuvenissen festival. All together, it was an interesting week, in which we learned a lot and had lots of fun.

FOR INFORMATION

Peter Dirix is now a researcher at the Centre for Computational Linguistics of the KU Leuven

Email: peedirix@hotmail.com

E-Masters Co-ordinator:
Frank van Eynde
KU Leuven – Centrum voor Computerlinguïstiek
Maria Theresiastraat 21
B-3000 Leuven
E-mail: frank@ccl.kuleuven.ac.be
Web: www.cstr.ed.ac.uk/Euromasters

Book announcement

New book announcement

Just published (published by Hermès in the IC2 series – Information, Command, Communication) in French:

Traitement automatique du langage parlé (Spoken Language Processing)
Edited by Joseph Mariani

Spoken language processing includes activities related to speech analysis; variable rate coding, in order to store or transmit speech; speech synthesis, especially from text; speech recognition and understanding, in order to transcribe it to text and eventually index it; and issues related to human-machine dialogue or human-human interaction with machine assistance. It also includes speaker and language recognition. Those various types of processing may be conducted in a noisy environment, which makes the problem even more difficult. These two volumes, dedicated to spoken language processing, address the following topics: how to realise speech production and perception; how to synthesise and understand speech, with the support of the presently available knowledge in signal processing, pattern recognition, stochastic modelling, computational linguistics, and human factor studies but also with the help of knowledge specifically related to speech.

In two volumes :
1- Speech analysis, synthesis and coding:
Speech analysis (C. d'Alessandro),
Speech coding (Gang Feng, L. Girin)
Speech synthesis (O. Boëffard, C. d'Alessandro)
Talking face synthesis (T. Guiard-Marigny)
Computational audio scene analysis (A. de Cheveigné)

2- Speech recognition:
Speech recognition principles (R. de Mori , B. Bigi)
Speech recognition systems (J.L. Gauvain, L.F. Lamel)
Speaker recognition (F. Bimbot)
Robust recognition methods (J.P. Haton)
Multimodal speech (J.L. Schwartz, P. Escudier, P. Teissier)
Speech in Human-Machine Communication (F. Néel, W. Minker)
Speech in telecommunications (C. Gagnoulet)

FOR INFORMATION

To send orders:
Lavoisier
14, rue de Provigny
F-94236 Cachan cedex
Tel. : + 33 (0)1 47 40 67 00
Fax : + 33 (0)1 47 40 67 02
Web: www.Lavoisier.fr

Volume 1:
ISBN : 2-7462-0440-1 • 60 Euros • 202 pages • 2002
Volume 2:
ISBN : 2-7462-0441-X • 75 Euros • 240 pages • 2002

What is ELSNET going to do in the next framework programme?

***Steven Krauwer**, ELSNET Co-ordinator*

ELSNET was created in 1991 with the objective of bridging two gaps: on the one hand the gap between the language and speech technology communities, and on the other the gap between academia and industry. Over the years we have built up a community of some 140 academic and industrial organisations active in language and speech technology (the members of the network), over 1000 other interested parties who are subscribed to our mailing lists and to *ELSNews*, and over 3500 experts and organisations listed in our directories of organisations and experts. Together with this community we have organised and supported numerous actions and events, such as our annual summer schools, training courses, conferences, workshops, and panels (often in conjunction with major conferences). We have published books and reports, and through our websites at www.elsnet.org and www.hltcentral.org (a joint enterprise with the EUROMAP project) we have collected and disseminated information relevant for our community.

All this activity would not have been possible without the funding provided by the EC, and without the active participation of a large number of members of our community. As most of you will know, the EC funding policy is based on projects, with a typical duration of two or three years. Over the years the members of the ELSNET team (the co-ordinator and his staff, the Executive Board, and the various task groups and committees) have (successfully) put in new ELSNET project proposals to ensure continuous funding for our community. We have managed to move seamlessly from one framework programme to the next from the very start, and we are now in the last phase of our funding from the EC's fifth framework programme (the ELSNET4 project, which started last summer and which will last until early 2004). ELSNET4 is not different from its predecessors except that the duration is rather short, and that our funding level is relatively modest. We have adapted our work programme accordingly by focussing on a few main topics, and we have simplified our management structure (the ELSNET4 project is now managed by a small management committee consisting of the core partners: Utrecht University (co-ordinator), DFKI, and University of Pisa. The traditional ELSNET Board (now consisting of some 14 prominent members of our community) has not disappeared and will remain active as the main body taking care of our community, but it will not be involved in the day-to-day management of the ELSNET4 project as such.

Where do we go from here? As I indicated above, we cannot preserve (let alone increase) our level of activity if we are not funded. Active involvement by our members is a necessary, but not a sufficient, condition to keep ELSNET alive as an active community. As the first call for proposals in the EC's new sixth framework programme (FP6) has been published, this is the right moment to start acting, and I would very much like to ask all the members of our community (not just the ELSNET members in the formal sense) to help us to prepare ourselves for the future, starting from the following two assumptions:

- ELSNET has over the years shown itself to be a useful instrument to facilitate and support R&D in the field of language and speech technology and deserves to be continued;
- In order to be able to serve our community optimally effectively and efficiently we should try to aim at doing more of the same, but also take into account that our environment has changed radically, not just in the bureaucratic and administrative sense (the working modalities in FP6 are very different from the past), and in the sense that the architecture of the programme has changed (FP6 is about big projects; no specfic programmes for language and speech technology), but even more so in the real world, where we see tendencies that we didn't see before. Just to mention a few examples: the movement of language and speech systems towards systems and applications that are embedded in other systems or work flows; the increasing linguistic complexity of the EU; the movement towards internationalisation and globalisation.

Our present default thinking is that the best way to preserve this community is to aim at a continuation of ELSNET as a so-called 'co-ordination action', where there is space for thematic networks dedicated to specific topic fields. We don't think that continuation of ELSNET as an FP6 style Network of Excellence would make much sense, because our scope is too wide and our membership too diverse to be able to set up joint research programmes. This should not of course prevent us from looking for possibilities to collaborate with such networks.

At this point I would like to invite all the members of our community to come up with ideas and ingredients for a programme of work for a new ELSNET project as a thematic network. This could include:

- things we would like to achieve

➡

- problems or topic areas we should address
- (types of) actions we should undertake
- things we should not do
- parties we should collaborate with or involve in the network
- offers to contribute actively to the new network and its creation

Please send me your ideas within the next few weeks. I will set up a page on the ELSNET website where I collect your reactions and where I try to keep you informed of the progress we make, the problems we encounter and the solutions we find.

Although I wouldn't want to change the name of the network (ELSNET has become a brand name), we need a name for the new project, and I propose to baptise it ELSNET6. This should reflect our connection with FP6 and at the same time do justice to the fact that the changes between FP5 (and its predecessors) and FP6 are dramatic enough to skip a number in the series.

FOR INFORMATION

The URL of the ELSNET6 website is:
http://www.elsnet.org/elsnet6

ELSNET roadmap on line!

***Steven Krauwer**, ELSNET Co-ordinator*

Two years ago ELSNET started creating a roadmap for human language technologies for the next decade.

A road map comprises an analysis of the present situation, a vision of where we want to be in ten years from now, and a number of intermediate milestones that would help in setting intermediate goals and in measuring our progress towards our goals. The function of the road map is not to impose anything on anyone, but rather to provide a broadly supported definition of a context in which to position the community's efforts, which would allow us to identify common priorities for joint activities in, e.g. research, resources, and training, and to detect major challenges and the interdependencies that may exist between them.

As a first step towards the roadmap we have organised seven roadmap workshops. Two of them were internal, with invited experts, and the other five were organised in conjunction with a number of major language and speech technology conferences, and were dedicated to various subareas of human language technology.

In close collaboration with DFKI we have now developed a graphical representation of our roadmap, inspired by the roadmap metaphor. Although not all workshop results have been incorporated in this presentation, the graphical version of the roadmap is now publicly accessible from our roadmap page. We would like to invite all members of our community to have a look at the roadmap in its present form, and to help us to construct – collectively – a broadly supported roadmap for our field as a whole.

The procedure we envisage, which is crucially dependent on your cooperation and contributions, is the following:

- The inclusion of results from past workshops will continue
- We will keep organising workshops to improve, extend and update the roadmap
- We are providing on-line facilities for people to react to the roadmap presented on the web, to express their agreement or disagreement, to elicit discussions, etc.
- As the world around us is continuously changing, we don't expect a roadmap to be valid forever – or even for more than a year. At this moment we expect to have a reasonably complete and stable version on the web by this summer, but we will continue to publish revisions at regular intervals, based on the feedback we collect via the web and at our workshops.

The internal representation mechanism we have chosen (an underlying database) allows for presentation in both graphical and in tabular format, and for easy updates and corrections.

FOR INFORMATION

A list of workshops (including programs, presentations and reports, if available) can be found on the ELSNET website at www.elsnet.org/roadmap.html

Calendar

Future Events

Feb 16-22 *Fourth International Conference on Intelligent Text Processing and Computational Linguistics*: Mexico City, Mexico
Email: gelbukh@CICLing.org URL: www.CICLing.org

Mar 20-21 *Semantics and Modelling Conference:* Paris, France
Email: jsm-meeting@erst.fr URL: semantique.free.fr

Mar 27 – Apr 1 *Corpus Linguistics 2003:* Lancaster, UK
Email: cl2003@comp.lancs.ac.uk URL: www.comp.lancs.ac.uk/ucrel/cl2003

Mar 31 – Apr 1 *Terminologie et Intelligence Artificielle 2003:* Strasbourg, France
Email: Rousselot@liia.u-strasbg.fr URL: u2.u-strasbg.fr/spiral/TAIA2003

Apr 1-3 *Voice World Europe:* London, UK
Email: stefan.nilsson@terrapinn.com URL: www.voice-world.com

Apr 12-17 *10th Conference of the European Chapter of the Association for Computational Linguistics:* Budapest, Hungary
Email: kindl@sztaki.hu URL: www.conferences.hu/EACL03

Apr 23-25 *Eighth International Workshop on Parsing Technologies:* Nancy, France
Email: Guy.Perrier@loria.fr URL: iwpt03.loria.fr

Upcoming submission deadlines

Feb 7 *TALN 2003*, Jun 11th-14th 2003, Batz-sur-Mer, France,
URL: www.sciences.univ-nantesfr/irin/taln2003

Feb 15 *HPSG-2003*, Jul 18th-20th 2003, East Lansing, Michigan, USA, URL: hpsg.stanford.edu/2003

Feb 26 *ACL03*, Jul 7th-12th 2003, Sapporo, Japan, URL: www.ec-inc.co.jp

Mar 1 *ISCA Workshop: Error handling in spoken dialogue systems*, Aug 28th - 31st 2003, Chateau-d'Oex, Vaud, Switzerland, URL: www.speech.kth.se/error

Mar 1 *MTT2003*, Jun 16th-18th, Paris, France, URL: www.mtt2003.linguist.jussieu.fr

Mar 31 *TSD 2003*, Sep 8th-11th 2003, Ceske Budejovice, Czech Republic, URL: www.kiv.zcu.cz/events/tsd2003

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.

Don't miss EACL!

Claire Gardent*, EACL*

The next conference of the European Chapter of the ACL, EACL03, is almost upon us. It will be held in Budapest, Hungary, from 12 – 17 April 2003.

The various calls for participations have been well heard resulting in very positive submission rates. We received 181 submissions for the main conference, 81 for the research notes session and 18 for the student research workshop. We also received 18 workshop, 14 tutorial and 20 demo proposals

After selection by the programme committees, the conference will thus feature 12 workshops, 4 tutorials and a very lively three days mixing invited talks, main paper, research notes, demos and student paper presentations.

This should make for an extremely interesting gathering. Don't miss it!
Looking forward to seeing you all in Budapest!

STOP PRESS!

The ELSNET Summer School will be held in July 2003 in Lille, France.

The topic is "Language and Speech Technology for Language Learning".

See the ELSNET web site for more details.

ELSNET

Office

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

Task Groups

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

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

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

Executive Board

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

The ELSNET Participants: Academic Sites

A	University of Vienna
A	Austrian Research Institute for Artificial Intelligence (ÖFAI)
A	Vienna University of Technology
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
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
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

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
DK	Tele Danmark
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
PL	Neurosoft Sp. z o.o.
RU	Russicon Company
RU	ANALIT Ltd
S	Sema Infodata
S	Telia Promotor AB
UK	Vocalis, Ltd.
UK	Imagination Technologies plc
UK	Hewlett-Packard Laboratories
UK	Canon Research Centre Europe Ltd
UK	ALPNET UK Limited
UK	Reuters Ltd
UK	SRI International
UK	Sharp Laboratories of Europe Ltd
UK	BT Adastral Park
UK	Logica Cambridge Ltd.
UK	20/20 Speech LTD

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. 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