elsnews 6.2

The Newsletter of the European Network in Language and Speech Apr 1997

Language Technology in Japan

Sophia Ananiadou, *Manchester Metropolitan University*

Japan's contribution to Language Technology has been impressive. Some of its achievements inevitably tie in with the properties of the Japanese language. Because of the phonological structure of Japanese, for instance, early successes were booked in the area of speech recognition and synthesis. Word-processing in Japanese, on the other hand, required the phonetic transcription of texts to be semi-automatically converted into Japanese characters (kanji), and posed extremely challenging problems to researchers. But here, too, the results have been remarkable. One of the biggest players in the Japanese NLP R&D private sector is **JUST**, which has the largest market share in word-processing software, and is Microsoft's strongest competitor in the Japanese market.

One important consequence of this success has been a favourable attitude towards LT technology within Japan, and generous funding for NLP research. Since the early '80s the Japanese Government has strongly supported research in speech and language technology, through projects such as **The Fifth Generation**, **EDR**, **Mu MT** and **Real World Computing**. Investment by major companies has been equally impressive, with the lion's share going to MT. At the same time, the private sector is still investigating how best to explore the potential of the most recent technological advances (such as Information Retrieval).

Generous funding thus constitutes one obvious area of advantage for Japanese Language Technology in comparison with Europe. Another important difference between the Japanese tradition and the European one is that research in LT in Japan is led by engineers, and has been application-oriented since the early '70s. This has resulted in a good relationship between research in the private and the public sector; but the downside is a degree of 'ad hoc-ness' and lack of abstraction.

Finally, the exchange of personnel between the private and the public sector is well-established in Japan. Engineers from the private sector often benefit from study leave at universities to acquire new skills and catch up on recent research results. This technology and personnel transfer between the two sectors has proven to be very productive; currently unique to Japan, it sets an example that Europe may well want to follow.

From Research to Application

Moving from research to product development lies at the heart of Language Technology, but contributions to this is issue of ELSNews suggest that it involves a complex balancing act. Theoretical and commercial considerations, the public and the private sector, user demands and constraints of cost and time all come into play, and need to be negotiated productively.

Which factors are crucial to the success of a Language Technology company? What can be learnt in the process of developing a commercial NL product? Should L&S research be more market driven? And — one of the most burning questions for some — is Europe lagging behind Japan and the US in the area of Language Technology?

Contributors to this issue of ***ELSNews*** *address these questions from various perspectives. We hope some of their answers and reflections will inspire you!*

The ELSNews Editorial Team

FOR INFORMATION

Sophia Ananiadou is Senior Lecturer at the Department of Computing, Manchester Metropolitan University. She has close contacts with Japanese Industry, and has collaborated with companies such as Matsushita since 1989. She recently spent a year in Japan as invited fellow at **NTT Communication Science Laboratories, EDR** (Electronic Dictionary Research Laboratories of Japan).

The author can be contacted at:
Department of Computing
John Dalton Building
Chester Street, Manchester M1 5GD, UK.
Tel.: +44 161 247 1542
Fax: +44 161 247 1483
Email: S.Ananiadou@doc.mmu.ac.uk

Apr 1997

elsnet

ISSN 1350-990X

Feature: Successful product development in Eastern Europe

From Research to Application...

Gábor Prószéky , MorphoLogic, Budapest

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

Editorial team: Mimo Caenepeel, Yvonne van Holsteijn and Steven Krauwer.

Lay-out and production: Mimo Caenepeel

ISSN 1350-990X

FOR INFORMATION

Contributions to ELSNews, and address corrections, should be sent to:

elsnews@let.ruu.nl

Fax: +44 131 650 6626

Material for the next issue is due:

15 May 1997

Commercial success in Language Technology is a relatively recent phenomenon. Between 1950 and 1980, the emphasis in the area was mainly on research; and it wasn't until the '80s that the proliferation of personal computers and major word processing applications increased the need for (simpler) language applications, and turned Language Technology into an industry. In Eastern Europe this process happened more slowly than in the West, partly because technology was less advanced, and partly because the formal structure of many Eastern European languages is more complex. In morphological terms, for instance, English, and even German and the Romance languages, are easier to describe than Slavic, Turkish or the Finno-Ugric languages.

But a disadvantageous situation, caused by the formal complexity of a language like Hungarian, can be turned into an advantage. This is what we have learnt at **MorphoLogic**, a Language Technology company founded in 1991. One of the original premises was that a computationally effective method can be theoretically sound, and, consequently, that a group of researchers with an academic background can be successful in the market place without having to give up high standards. Our experience in recent years has confirmed this, and, even though it is inevitably limited, it may be useful for other people who want to move from academic research to marketable industrial applications.

Proofing tools are among the best-known Language Technology applications used for word processing. Their spell-checkers and hyphenators are usually based on simple monolingual sets of words, (like wordlists and thesauri), and their grammar checking methods on heuristic pattern matching. But such a non-linguistic, wordlist-based approach is no longer a viable solution for more recent, or emerging, office and business applications of Language Technology (like computer-assisted language learning; intelligent data base indexing; "noiseless" free text search; intelligent automatic extraction and document indexing; intelligent 'find and replace'; automatic selection and categorization of faxes and emails by 'reading' them; handwriting and speech recognition; correct segmentation of non-segmented input; and synchronized handling of different language versions of translated documents). Relying on different linguistic and/or software strategies, on the other hand, is not the answer either: their integration into office and business automation systems would cause useless multiplication of very similar resources. What is needed are high-level language engineering programmes.

In response to this, **MorphoLogic** has developed its own linguistic software technology, as well as linguistically sound algorithms and data with computationally effective implementations, without typical tricks and ad hoc solutions. In the process, we have found that

- applications should be as language-independent as possible: multilinguality is important both in terms of research and in terms of prospective marketable applications.
- a modular, portable, platform-independent, and application-independent approach works best, not only for wordprocessor-based applications, but also for a wider application area.
- a balance needs to be struck between fully automatic methods and methods that rely on user interactions.

Although managers of small LT enterprises are usually research-oriented, we would hold that an industrial LT firm needs a separate sales/business manager to develop a profit-making enterprise. Particularly in former Eastern bloc countries, such a manager needs to be aware of both political and economic factors. For instance, the current political situation is moving towards Eastern European countries joining the EU; but at the same time a change in direction, although unlikely, is possible. It is therefore important for Eastern European LT firms to develop exports and non-local markets to some degree, but without coming to depend on them. Similarly, a sensitivity to economical factors — CEE countries are still in a recession, but there are signs of an economic revival — is necessary.

In addition to this, it is important for people running an LT company to understand the general climate of the industry they work in. This means, for instance

- keeping a balance between promises and reality in advertising (it is tempting to advertise the company's "super technology" as "solving everything")
- protecting the home market: being successful on the home market is often a prerequisite for international success.
- developing good relations with local linguistics experts, business partners and major purchasers, both locally and in other countries. It is the partnership with big software firms, which incorporate the technology from the LT firm in question, which will generate cash: if they turn away from the company, it could run into financial trouble.
- finding a balance between income generated by products and income from academic grants. The former keeps the company alive, the latter sparks the interest of employees with academic background.
- avoiding staff losses: NLP is a highly specialized field, and it takes time to train a new specialist.

To come back to **Morphologic**: since its start in 1991 no researchers have left our group. I believe this is due to both its scientific success and the relatively stable business situation: in combination, these guarantee new types of LT developments that are hard to find anywhere else in our country.

FOR INFORMATION
The author can be contacted at:
MorphoLogic
Németvölgyi út 25
Budapest
H-1126, Hungary
Tel: +36 (1) 201 8355
Fax: +36 (1) 155 7 155
Email: proszeky@morphologic.hu
WWW: http://www.morphologic.hu

... and from Networks to Lattices

John McNaught, *UMIST*

When I agreed to contribute an article to ELSNews, the following 'suggestions' of topics duly appeared on the back of an e-mail envelope, followed by an Editor's scrawl that looked rather like 'or ELS...':

- *Suppose you were a European minister or official in charge of Technology Transfer, and you had an unlimited budget at your disposal; could you describe three major steps, or initiatives, you would take to tackle Technology Transfer efficiently and productively?*
- *To what extent do these measures correspond to, or differ from, the kinds of activities that ELSNET is developing? Are any changes or adjustments called for, in your opinion?*

Now, technology transfer (TT) is a Big Topic, so I chose to concentrate on TT in language technology. Even then, I knew very little about this topic: after all, academics are hardly the best people to comment on TT, especially those from systems that reward the pursuit of research or teaching excellence and thus implicitly de-emphasise technology transfer activities. However, that thought gave me an idea for a first initiative: Let us create a European network of language technology transfer and information centres (LATTICE!), certainly down to regional level.

This is not wholly a new idea. Indeed, there are already a few groups and consortia that are active in TT in our domain (e.g. **CST** in Copenhagen, **LTG** in Edinburgh, and the **Verbmobil** Consortium). However, these operate independently, when outside ad hoc consortia. My thinking is that TT can only really take place successfully if people are a) aware of the technology and its applications and b) able to come together to work on the specification, construction or evaluation of a practical application. It is in working together that true, lasting, effective transfer comes. Better transfer is achieved in small dedicated groups working in the same regional location, as opposed to amorphous internationally distributed project consortia. If industry (both users and developers) is to benefit from TT from academia, and vice versa, people must be able to inspect and evaluate the current technology, state of the art research protoypes, and so on; to learn about upcoming developments; and to discuss their needs with various experts (including business and management experts). That is, they need information and hands-on experience. TT centres could certainly act as regional showcases and advice centres for LT.

Currently, much effort is being put into the information gathering and dissemination part of language technology TT (e.g. **LE EUROMAP**). There is a nascent network of Innovation Relay Centres (IRCs) — established under the EC's Innovation Programme— which offer general TT information, advice and contacts, and organise seminars and training. These are already valuable services. ELSNET is of course very active in this area too. However, my conception of LATTICE involves TT centres dedicated to LT. This is motivated by what **EUROMAP** calls the pervasiveness of LT, weak national support for LT and the lack of understanding in the business community. It also involves the notion of a diverse core of language engineers in each TT centre, working either directly on industrialisation aspects, or on facilitating collaboration with specialised SMEs, for example. The TT centres could also become engaged in resource provision and/or customisation (running 'grammar factories', providing domain models). A core team in each centre would provide stability and continuity of expertise — something that is missing in our rather fluid, short-lived, project-dependent research teams.

Given the multilingual nature of LT, the increasing integration of speech and NL in applications, and the need for large-scale reusable resources, a network is called for: we must cooperate to succeed, both among TT centres and with organisations such as **ELRA**. I note, in passing, the central role of LT standards in underpinning the spread and success of TT centres. Indeed, initiatives such as **EAGLES** (the most successful of the **LRE** projects and still going strong) would have even greater impact through industrial participation via TT centres.

Finally, we are constantly told that there are not enough team leaders, not enough engineers and not enough students in language engineering. Partial causes of this are the lack of job prospects and the lack of LT degree programmes at undergraduate level. The existence of TT centres could serve a secondary but vital purpose in contributing to student training, gaining of industrial experience, and preparation for the job market, as well as creating a greater need for LT-related jobs in industry. In this context, let me briefly propose two other initiatives, both predicated on the existence of LATTICE: the creation of a Language Technology Transfer Association; and the creation of a Language Technology Education initiative, to ensure that educational programmes over the entire educational spectrum address LT issues in some form or other (the latter in cooperation with my colleague holding the Education portfolio).

Regarding the second question, briefly, many of ELSNET's activities are well-oriented to information, dissemination and training, and it is clear that these would be complementary and valuable to any future LATTICE. There would be a useful role for ELSNET Managing and Associate nodes in collaborating with regional TT centres, or helping to set them up in the first place. I would urge early cooperation between ELSNET and the IRC network, to promote the cause of LT.

FOR INFORMATION
John McNaught is Co-Chief Editor of EAGLES. He can be contacted at:
Centre for Computational Linguistics
Department of Language Engineering UMIST
PO Box 88 Sackville Street
Manchester M60 1QD, UK
Email: jock@ccl.umist.ac.uk
Tel: +44 161 200 3098
Fax: +44 161 200 3099

Kielikone Machine Translation: How We Did It

Harri Arnola, Kaarina Hyvönen & Tim Linnanvirta, Kielikone Ltd.

The road from a 'brilliant idea' to an end user product can be long and arduous, but, as the Finnish company ***Kielikone Ltd.*** *found out, it can be worth travelling.* ***Kielikone Ltd.*** *developed a successful commercial product, the MT Translation system* ***TranSmart. Harri Arnola, Kaarina Hyvönen*** *and* ***Tim Linnanvirta*** *share some insights in what was learnt in the process.*

Kielikone Ltd. is a Finnish language technology company that has produced both generic machine translation technology and an MT system based on that technology. What started out as a revolutionary idea has since evolved into a successful product.

Kielikone's roots trace back to the '80s, when the **Sitra Foundation** funded the **Kielikone project** to study computational aspects of Finnish. In 1987 the Kielikone project was approached by a number of industrial partners with a proposal for building a machine translation system, to alleviate the great burden of translating technical documentation. The result was a machine translation project, funded first by the Sitra Foundation and later by the Technology Development Centre of the Finnish Ministry of Trade and Industry. The industrial partners formed a consortium for the project and shared the costs. **Kielikone MT** was born.

It can now be stated proudly that Kielikone MT is a success. Its MT system, **TranSmart,** translates from Finnish into English. It runs on various Unix machines as well as on Windows NT (under Windows NT, on a 200 MHz Pentium, the translation speed is about 15 sentences per second), and is in production use in several companies. It also continues to be improved: translation rules and lexical entries are refined when necessary, and new lexical entries are added whenever faults are found. In November 1996, **TranSmart** received a prestigious award as the best Finnish software product of the year.

Kielikone has also established a successful machine translation services company, **Transwise Oy,** in collaboration with the translation company **Trantex Oy**, and operates in other product areas as well. It is, for example, the market leader in the field of electronic dictionary products and proofreading software in Finland. But in this article we concentrate on machine translation R&D work. Through a set of 'dicta' — some of which may be obvious, others perhaps controversial — we hope to convey (some of) the experience we have gained in the strenuous journey to where we are now.

Be humble!

Anyone thinking seriously of building an MT system should be (and usually is) fully aware of the fact that any MT system will be markedly less competent than a professional human translator. If we rank professional human translators at the top of a scale as 'complete' translators, MT systems will always remain incomplete in comparison.

But that does not mean that MT work is worthless. Once upon a time **Bar-Hillel** claimed that machine translation was impossible because no system could properly disambiguate the two meanings of the word *pen* in the sentence *The ball was in the pen.* All such an argument actually shows, however, is that there are sentences which, for semantic reasons, cannot be properly translated automatically using a single, general translation lexicon. Bar-Hillel's statement says nothing about how far MT can go, statistically speaking, or about its practical value.

Due to this inherent incompleteness, MT research needs to progress through prototyping. A project should aim for an early prototype that gives statistically reliable data about the quality and efficiency of the planned approach. And one should be prepared to adjust the approach, or even discontinue the project, if that is what an evaluation of the prototype suggests.

Be theoretical!

The word 'theory' means different things to different people. In linguistics, for instance, a system or framework is often not considered theoretical unless it complies with the pet theory of the judge. But in this dictum 'theory' refers to a consistent set of generalizations about linguistic structures.

Translation is about associating words and strings of words from one language with their counterparts in another language. Little is known about the processes in the mind of a competent human translator when he or she makes such associations, except perhaps for the vague statement that there is probably some deep semantic processing involved. If MT were not theoretical, it would end up directly associating words and strings of words with their counterparts in the target language. Such work would progress slowly, and either never even appprach completion, or have qualitatively poor results.

To be practical, MT has to associate structures with structures, and since structures are theoretical constructs, MT must be theoretical. **TranSmart** performs a full morphological analysis of word forms and a full syntactic analysis of sentence structures before translation takes place.

Stick with it!

This dictum probably applies to any software project which is expected to take years or decades to finish. Sooner or later some bright young man or woman will join and explain that the project is worthless unless it utilizes a new methodology, say, neural networks. Then there will be somebody else with yet another new idea. If you agree to redesign the fundamental principles of the project every time such criticism arises, the project will never reach a product state.

So choose your fundamental paradigm wisely, and stick with it!

Be thrifty!

There are various alternative linguistic theories available. Engineering is appropriately called 'the science of scarce resources', and MT belongs squarely within language engineering. In choosing between theories, MT should give heed to the ones that are least wasteful.

All or most MT systems run on serial processors, and in serial processing time is a scarce resource. One should beware of taking a fancy to a linguistic theory that purports to reveal something important about human linguistic processing, and pay attention instead to theories that yield efficient computation in serial processors.

We do not believe that theories of the constituent structure of sentences are pertinent from the point of view of machine translation. We would hold instead that functional descriptions of sentence structure are more apt in MT, since functional structures are close to logical structures. Our own choice has been Dependency Theory. Our parser demonstrates that it is possible to parse dependency structures of at least one language in linear fashion.

Ignore the siren calls of semantics!

It has been a refreshing pastime for linguists to find examples that cannot be translated by a machine without resorting to semantic processing, and to conclude that MT needs sophisticated semantic processing. We already cited **Bar-Hillel**'s contribution in this respect. Other examples in the same vein are not hard to find. We could easily come up with several Finnish sentences whose proper translation into English requires semantic consideration.

However, in everyday life we do not expect tools and utensils to apply in all possible circumstances. They are usually designed for limited application only, and many possible attributes are left out because their implementation would be too costly compared to the potential benefits. The same applies to MT. Deep semantic processing is costly for MT in two major respects. First, there is a great amount of human intellectual labour involved in associating words with sophisticated conceptual structures. And secondly, the processing of such semantic information would represent considerable computational overhead.

We deliberately left advanced semantic processing out of the first version of **TranSmart**, and decided instead to study how far we could go using dependency structures only. **TranSmart** employs shallow semantic processing: words are classified in semantic typologies, and these semantic markings are used as selection restrictions. Hindsight tells us that this was a wise decision.

Part of the semantic problem can also be solved by a simple and widely applied strategy: the use of multiple lexicons. **TranSmart** is able to access several consecutive translation lexicons: first a user-specific lexicon, then a domain-specific lexicon, and finally a general lexicon.

Remember what happened to Napoleon!

Napoleon tried to conquer Europe and ended up being locked in a house on an island, with all his ambitions shattered. MT research should not try to conquer all the languages of Europe in one sweep, using a great army of researchers. Prudent research focuses on a single language pair, and employs a small and tight group of bright people. If the language pair comprises two structurally distant languages, the chosen MT approach may prove applicable to a wide range of other language pairs later on.

Kielikone focuses on the Finnish-English language pair. The R&D group contains an average of 5 persons, all working within shouting distance from one another.

Remember that great invention: the ladder!

The ladder was a great invention. Despite it being such a simple device, it makes it possible for a person to reach extraordinary heights in small consecutive steps. Translating from one language into another is a difficult task. But by emulating the ladder, MT research can progress through small but stable steps, and reduce the complexity of the whole problem into manageable chunks.

Kielikone uses the ladder strategy both in the design process and in the system architecture. The design process has produced generic linguistic processors which serve as stable stepping stones. Morphological analysis and dependency analysis produce general results which are useful in other contexts as well. In terms of its system architecture, **TranSmart** consists of a sequence of generic MT engine applications, each performing a certain linguistically motivated subtask. Each phase adds new information without destroying old data. Such an architecture is flexible and allows for experiments with various ways of decomposing the translation process into subtasks. A generic MT development environment was built before a specific MT system saw the light of day — yet another application of the ladder strategy

Remember that the customer (or user) is king!

Researchers on any software development project should carefully listen to the user, since a software product can only rarely be imposed on a user against his or her will. This consideration is particularly important for MT, since MT projects are always expensive and time-consuming. It is wise to listen to the users at an early stage. Kielikone MT had the support of a user consortium from the very beginning of the project, and our success can be attributed at least partly to the feedback given by the beta testers.

FOR INFORMATION

The generic MT environment (**KMTech**) developed by Kielikone MT R&D can be used to design MT systems for various language pairs. Kielikone offers KMTech for joint projects on the construction of MT systems. An interactive demonstration of the TranSmart Finnish/English MT system and the DCParser can be found on the Kielikone homepage.

The authors can be contacted at:
Kielikone Ltd.
P.O. Box 126
00211 Helsinki,
Finland
Tel: +358 9 6820 211
Fax +358 9 6820 167
Email: harri@kielikone.fi
WWW: http://www.kielikone.fi

Current Developments in Natural Language Processing at Apple

Branimir Boguraev, Apple Computer, Inc. & Christopher Kennedy, University of California, Santa Cruz

Over the past several years, Apple Computer, Inc. has been actively pursuing the goal of developing its natural language program — both in terms of establishing a core set of NLP technologies, and in defining application areas for these. This development is under the direction of Dr. **Branimir Boguraev**, formerly of IBM's T.J. Watson Research Centre and the Computer Laboratory at Cambridge University. Organizationally, the natural language work is carried out at the Apple Research Laboratories; it is positioned within the Intelligent Systems Program (Dr. **James Miller**) in the Knowledge Management Group, and is coordinated with some on-going speech work (Dr. **Kim Silverman**) and information access research (Dr. **Dan Rose**).

Projects under **Boguraev**'s supervision investigate a range of issues, including:

- optimal packaging of a substrate of NLP functionalities, with appropriate Application Programming Interfaces (API's), embedded within the Macintosh Operating System (Mac OS);
- their synergistic integration with other information technologies;
- studies of how NLP can be leveraged for further enhancing the user experience; *and*
- building several information management systems incorporating linguistic processing of text-based documents.

Under Apple's internship program, **Boguraev** works with graduate students in Linguistics and Computer Science, assisting in the development of prototype systems; Apple also participates, jointly with the Computer Science Department at Brandeis University, in NSF-sponsored projects within the Human Language Technology program, looking at some more practical aspects of current research in computational lexical semantics. Especially within Apple, the emphasis has been on finding suitable tasks within which to embed linguistic functionalities; on striking the right balance of scalable and robust technologies which can reliably analyze significantly large text sources; and on developing algorithms for focused semantic analysis starting from a relatively shallow syntactic base.

A case in point is technical terminology identification. Traditionally, this has had somewhat limited use, primarily for indexing purposes; at Apple, however, it has been applied, within a general domain acquisition framework, to the task of instantiating databases for instructional assistance. In particular, an NLP environment has been customised to derive help databases automatically, by parsing on-line software manuals. This work was largely carried out as a joint effort with **Michael Johnston** (previously with the Linguistics Department at the University of California at Santa Cruz, and now at the Oregon Graduate Institute) and **Scott Waterman** (formerly of Brandeis University, now with Price Waterhouse Research Laboratory). A more recent project brought together **Chris Kennedy** (Linguistics, University of California at Santa Cruz) and **Marc Verhagen** (Computer Science, Brandeis University); the focus here is on semantically-driven content analysis of arbitrary texts, and exemplifies certain aspects of the more algorithmically-oriented work at Apple.

On-line help and domain functionality

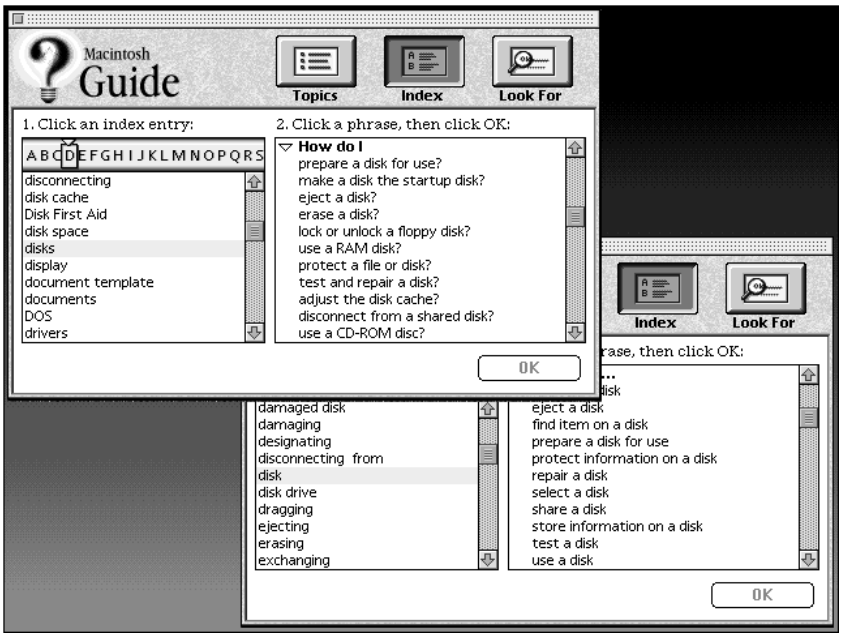
Apple Guide is an integral component of the Macintosh operating system; it is a general framework for on-line delivery of context-sensitive, task-specific assistance across the entire range of software applications running under the Mac OS. The underlying metaphor is that of answering user questions like *"What is X?"*", *"How do I do Y?", "Why doesn't Z work?" or "If I want to know more about W, what else should I learn?"*. Answers are 'pre-compiled', on the basis of a full domain description defining the functionality of a given application. For each application, assuming the existence of such a description in a certain database format, **Apple Guide** coaches the user through a sequence of definition panels (describing key domain concepts), action steps (unfolding the correct sequence of actions required to perform a task or achieve a goal), or cross-reference information (revealing additional relevant data concerning the original user query).

An **Apple Guide** database is typically instantiated by hand, on a per application basis, by trained instructional designers. Viewed abstractly, however, the information in such a database constitutes complete domain characterization for the application—in terms of domain objects, their properties, and relations among them. In order to answer questions like those above, certain aspects of the domain (in this case that of operating system level activities) need to be identified: a kind of a domain object is a disk; there are several types of disk (including floppy disks, start-up disks, and internal hard disks); disks need to be prepared for use; floppy disks can be ejected; and so forth.

It is clear that a terminology identification component, applied to suitably chosen technical documentation for the domain, could be profitably utilized for the purposes of such domain characterization. We have found that the core set of terms from a technical document can be refined not only to include all (and only) the domain objects, but also to enrich the descriptions of these domain objects, by deriving relational structures for each of them (not unlike generating lexical entries for a dynamically induced lexicon). This type of staged lexical acquisition ultimately derives a conceptual map of the technical domain. Mapping from such a domain description to an **Apple Guide** database is relatively straightforward. Terms are 'place-holders' for definitions of salient domain objects. Relations naturally map onto a *"How do I ...?"* panel. In database format terms, this would mean definition

entries for *"start-up disk"*, *"network connection"*, *"System folder"*; and action sequence panels for *"How do I specify a start-up disk?"*, *"How do I use the internal hard disk as a start-up disk?"*, and so forth. The definitions and task sequences would still have to be supplied externally, but the generation of the database is fully automatic.

The outcome of such a domain acquisition and mapping process is illustrated below.



One of the screen snapshots is from the 'canonical' Macintosh Guide for the standard Mac OS configuration; the Guide database here is built, manually, by a team of instructional designers. The other snapshot displays, through the same delivery mechanism, a database which has been fully automatically constructed, by the system outlined above, following an analysis of the primary technical documentation for this domain (Macintosh User's Guide). Note, in particular, the *"How do I..."* leading to detailed instructions concerning common tasks with specific objects (in this example, disks) in the MacOS domain. Barring nonessential differences, there is a strong overlap between the two lists (*"prepare a disk for use"*, *"eject a disk"*, *"test (and repair) a disk"*, *"protect a file/information on disk"*, and so forth). Moreover, some additional action types have been identified, which are clearly relevant to this domain, but missing from the 'canonical' database: *"share a disk"*, *"find items on a disk"*.

Content analysis and document characterisation

More recent work addresses the problem of identifying the core content-bearing units in arbitrary texts, with a focus on smaller documents (in comparison with the technical prose discussed in the previous section), and allowing for wide diversity of genre. This work is centred around the development of a set of sophisticated text processing tools based on a very shallow linguistic analysis of the input stream, in which depth of base level analysis is traded off for breadth of coverage (the analysis engine currently used is **Lingsoft**'s supertagger). Two complementary lines of attack here are exemplified by **Verhagen**'s work on identification, extraction, classification and typing of proper names, technical terms, and other complex nominals from text, which extends the core phrasal analysis engine originally developed for term analysis, and **Kennedy**'s work on algorithms for topic-based segmentation of text, salience-driven anaphora resolution, and content characterisation.

The work on anaphora resolution provides a representative illustration of our goal of deriving high-level semantic analysis from an impoverished input stream. The basic strategy we employ is a modified implementation of an algorithm developed by **Lappin and Leass**, which relies heavily on fully parsed inputs. We have found that by combining the phrasal analysis generated by an extended implementation of the term identification technology with an analysis of the overall topical structure of the text (derived by comparing adjacent blocks of text for overall lexical similarity, after **Hearst**), we can achieve precision of resolution comparable to that of Lappin and Leass' algorithm. Our algorithm, together with its implementation and detailed evaluation, is presented in a COLING-96 paper.

Like the Lappin and Leass algorithm, our anaphora resolution procedure determines the *salience* of all referential expressions in a text. Roughly speaking, salience is a measure of the relative prominence of an object in a discourse: objects with high salience are the focus of attention; those with low salience are at the periphery. This measure is not only an important factor in determining the antecedent of a pronoun; it also provides a basis for establishing a partial ordering on a term set, which may then be used as the basis for a characterisation of a document's content in terms of those expressions which identify the most prominent participants in the discourse. Ongoing work at Apple is aimed at using this reduced set of terms (in combination with relational information of the sort discussed in the previous section and packaged in an appropriate presentation metaphor) as the basis for a highly representative and meaningful—but at the same time compact—document abstraction.

FOR INFORMATION
Branimir Boguraev can be contacted at:
Apple Research Laboratories, Apple Computer, Inc.
One Infinite Loop, MS: 301-3S
Cupertino CA 95014, USA
Tel: +1 408 974 1048
Fax: +1 408 974 8414
Email: bkb@research.apple.com

Christopher Kennedy can be contacted at:
Department of Linguistics, Stevenson College
University of California at Santa Cruz
Santa Cruz CA 95064, USA
Tel: +1 408 459 4765
Fax: +1 408 459 3334
Email: kennedy@ling.ucsc.edu

The Use of Corpora in Building Speech Applications

Harald Höge, Siemens AG

What is the role of corpora in application building? Should corpora used in the development of speech applications share special design features, or should they be adjusted depending on the individual type of application? What can data collectors do to adapt corpora to the needs of application building? ***Harald Höge*** *addresses these questions in the second part of our series on the use of corpora in NL & Speech.*

Recent trends in speech technology

In recent years, speech technology has started to play an increasingly important role in the building of commercial applications. Speech recognition, speech synthesis and speaker verification systems embedded in speech dialogue systems have all reached the market, and some of them are already proving to be very successful. The equipment value of telephone-based speech dialogue systems (or *Interactive Voice Response Systems*), for instance, reached a volume of $1.000.000.000 (1 billion) in 1996. Another system of commercial relevance is the 'hearing typewriter', which has found its first application in medical reporting. Efforts to broaden the applications for dictation are underway, and on the basis of current rapid growth in the computing power of PCs, it is estimated that unrestricted speech dictation will be a standard product for PCs in 3 - 5 years' time.

There is, of course, much room for development and growth. Thus, the first speech recognition systems for command input on PCs are available, but they are not very robust yet (and they have to compete with the mouse as input device). Speech Synthesis still suffers from poor speech quality, and is used only for very specific purposes. On the other hand, statistically-based methods for improving speech quality are currently being developed, and may soon bear fruit. And finally speech translation, although in its infancy, is emerging as a core technology area for the next century.

Given these trends, much effort is invested these days in increasing the number of applications based on speech technology. The main challenge lies in achieving the following goals:

- *robustness* (performance should not deteriorate in the concrete application environment);
- *multilinguality* (creating a new application for a new language should be easy); and
- *configurability* (it should be possible to integrate the speech device/software into an application at low cost).

Application-Specific Spoken Language Resources (ASSLRs)

To achieve these goals, two types of resources are needed: corpora and lexica. Statistically-based methods, whereby model parameters of a speech device are trained on large corpora, have proven to be very successful: the device effectively learns by example. But corpora do not contain all knowledge needed for an application, and linguistic and application-specific knowledge also have to be incorporated into the speech device. This is done mainly via lexica.

Corpora and lexica used for designing speech application are called *Application-Specific Spoken Language Resources (ASSLRs)*. The following are examples of ASSLRs:

- annotated speech databases
- text databases
- pronunciation lexica
- lexica with syntactic and semantic features
- lexica for abbreviations and acronyms

Unlike other Spoken Language Resources (SLRs) used in research, ASSLRs have to meet the criteria of robustness, multilinguality and designability. Best results are achieved when the ASSLRs are generated in the application environment. This is because of the statistical methods used. The statistical parameters of a speech device reflect only the statistical properties of the corpus used for training. If the corpus deviates from the application this will usually cause a degradation in performance. Some examples:

- The performance of a speech recognizer trained on speech databases created under 'office' conditions will degrade drastically if it is run in a mobile telephone environment.
- For telephone applications activated in a car using GSM transmission, the ASSLRs will have to cover car noise, GSM channel characteristics and Lombard effects.
- Regional dialects may adversely affect the performance of speech recognizers. For example, the ASSLRs created for the Spanish language in Spain will have to be adapted for use in Chile (a new speech database will have to be collected, and the pronunciation lexicon will have to be changed).
- In a linguistic lexicon, different semantic uses of the same word in different dialogue systems have to be reflected in corresponding semantic entries.

Most major companies involved in speech technology (e.g. **AT&T**, **IBM**, **Microsoft**, **VCS**, **Philips**, **Siemens**) are currently in the process of building ASSLRs. As the creation of ASSLRs is costly and time-consuming, consortia and distribution channels have been formed on a collaborative basis. They include:

- **SpeechDat**: a European project set up to produce annotated speech databases for telephone applications. Covers all EU languages and their major dialects.
- **ONOMASTICA**: a European project creating pronunciation lexica for names, covering all EU languages.

- **LDC** (an institution founded by DARPA): distribution channel for research SLRs and some ASSLRs
- **ELRA**: distribution channel for research SLRs and ASSLRs

Problems

Problems remain in two major areas. The first one concerns quality assurance of ASSLRs: there is a lack of standards with respect to ASSLRs, and the quality of current ASSLRs, in terms of mistakes made in annotating speech and text databases and encoding lexicons, is insufficient. This results in high costs in reusing ASSLRs for new applications: in our experience, making ASSLRs usable for an adapted recognizer takes several weeks.

Attempts are being made to address these issues. Within the **SpeechDat** project, standards have been set for speech databases designed for telephone applications. And in addition to this, a validation centre (**SPEX**) has been set up to assure ASSLR quality. But a lot of work remains to be done. Standards need to be set for all relevant ASSLRs, and ASSLRs available via distribution channels need to be validated.

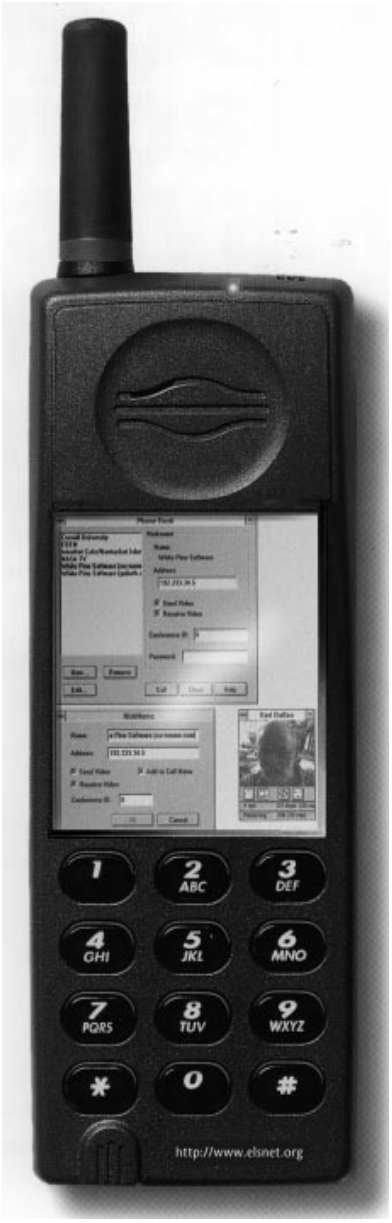
A second problem area concerns coverage in terms of applications and languages. The problem of language and dialect coverage is particularly evident in Europe. Coverage of applications relates to:

- channel characteristics (microphones, transmission channels, room characteristics),
- noise characteristics (office, car, public places, industrial environment), and
- terminology (lexica for specific applications, e.g. medical, legal, technical domains).

To create ASSLRs for each application separately is not always feasible, due to constraints of cost and time. So the general approach is to use more or less generic ASSLRs, which cover an area of applications, and to tune these to the specific application. In such cases a compromise has to be made between the costs involved in adapting the ASSLRs on the one hand, and the deterioration of the functionality of the speech device if the adaptation is less than perfect on the other.

Ideally, all speech devices should be easily adaptable to new applications using generic ASSLRs, without degradation in performance. A lot of research is currently being devoted to this. Here are some examples:

- in-service adaptation of dialogue and dictation machines: adapting HMM-Models (statistical phoneme models), language models (statistical grammar) and lexica during the use of the recognizer
- application modelling: adapting the parameters of the speech device off-line
- multilingual adaptation: avoiding multiple modelling of things which are almost identical across languages



But these techniques are only partly applicable for building speech devices in the next 2 or 3 years: ASSLRs covering a wide area of applications are needed. In terms of coverage, the two generic application fields — telephone and office applications — should be covered.

Conclusion

The increasing commercial relevance of speech technology makes the construction of suitable ASSLRs crucial. The most urgent steps to be taken in this respect are:

- creating multilingual SLRs covering the two generic application fields (telephone and office applications),
- setting standards and establishing procedures for assuring ASSLR quality, and
- strengthening research into adaptation methods.

FOR INFORMATION
Harald Höge is Head of the Speech-Group of Central Technology of Siemens Cooperation, and can be contacted at :
Siemens AG
Otto-Hahn-Ring 6
D - 81730 München
Germany
Tel: +49 89 636 53374
Fax: +49 89 636 49802
Email: harald.hoege@mchp.siemens.de

Is Language Technology coming of age?

ELSNews talked to Marc Moens, manager of the Language Technology Group at the University of Edinburgh.

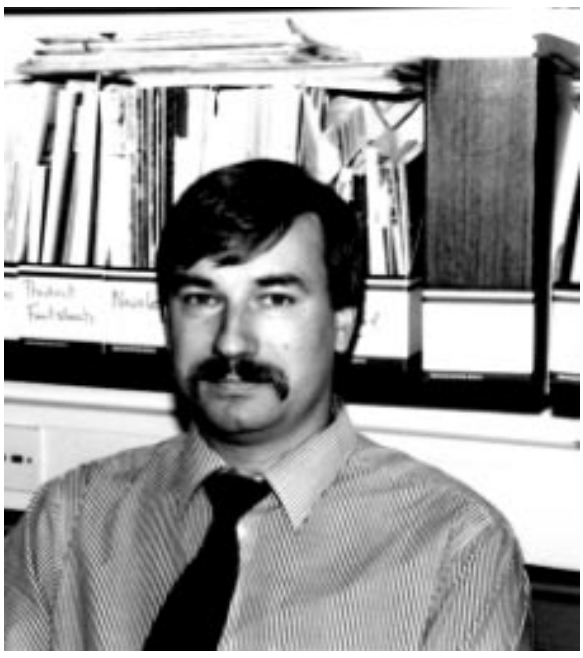
ELSNews: *Language technology seems to have come of age recently, with articles in scientific journals and advertisements for language products in the newspapers. Has there been a major breakthrough?*

Moens: It's certainly true that speech technology has reached the stage where major companies have found an angle to start marketing it. But that doesn't mean there has been a major breakthrough in speech technology. A few weeks ago, I bought a voice dictation system for my PC. It's by no means perfect. When I read in an existing text it gets nearly everything right. But when I try to dictate a message that I'm still composing in my head (a much more common scenario) I find it much harder to ensure that there are short pauses between the words, and the system gets confused. However, I can imagine that it is selling well. The version I bought only costs about £90. At such a low price, users won't be too disappointed when they discover its limitations.

So I think the breakthrough is from a marketing point of view. Once lots of people have a little dictation system on their PC, they'll be ready for more advanced and more expensive systems: systems that take continuous speech, that don't need a head-mounted microphone, that have a wide vocabulary, and so on. There is still a lot of research to be done before we get to such systems. But the big companies may be creating a market demand for them (and for the research) by marketing the simpler systems.

ELSNews: *It seems to be American companies doing this. Is Europe lagging behind in producing language and speech products?*

Moens: It's too early to say. American companies just have a very prominent presence in all aspects of IT. And they do have the advantage that they can mass market English language products for their home market. But the phenomenon I described with voice dictation systems has been happening in Europe with machine translation: you can get rough and ready translations over the Web, for example. When you find a Web page that you think you might be interested in, but it's in a language you don't understand, you can send it off and have it translated. That is a good way of introducing the power of machine translation into many people's daily work practice. And once people get used to that kind of service and realise that they do make regular use of it, they will want to have that kind of service more readily available, for more languages, just as fast but perhaps higher quality, and so on. At that point there will be a real market demand for that type of system, and for the research.



Marc Moens at the Human Communication Research Centre

So it's not so much American companies that are leading the way, it's more a case of big companies, that can afford to market less than perfect language technology at a cheap price, or for free. Text summarisation is another area where there has been a similar trend: **BT** offers free summarisation of Web pages; **Microsoft** has included a document summarisation feature in its Office 97 package; **Oracle** has a text summarisation feature in its document management software — they are all large companies. This summarisation software still has a lot of rough edges to it — indeed, many people would say that what the software does is not really "summarisation". But the software performs a useful function, and once people get used to it, they'll want more and better and faster summarisation. Which opens up a whole area of research.

ELSNews: *Does this mean that language and speech research should be more market driven?*

Moens: Not all of it. There has been a tendency in some European funding programmes the past few years to concentrate on language and speech research that is close to market, that has, for example, users involved from the start. You then get the problem that solutions are not generic: you identify a very concrete product or service, set up a consortium with users, and immediately start building a prototype, since otherwise the users won't have anything to do in the project. What is ignored is an analysis of how general the problem is and what the generic solutions to the problem are. When too much R&D is close to market, there is a danger that one starts producing lots of little prototypes, with no clear understanding of how the prototypes are related, how they scale up — i.e. very few transferable results. It's often those transferable, generic results which are needed to turn products into better products.

ELSNews: *So how should we see the relation between R&D and product development?*

Moens: Industry and academia will continue to make major breakthroughs in all areas of computing, including language and speech. What is needed for those breakthroughs are skilled R&D people. Universities should be the ones delivering these skilled people. One of the best ways of delivering people with the right skills is by educating them in an environment where they carry out combinations of basic and applied research. And

for a university department to be able to offer its students such an environment, it needs a thriving research programme. That should be the point of public funding of R&D: to ensure that public research labs continue to do good basic or applied or problem-oriented (or whatever) research, so that they can continue to turn out highly skilled workers. Their goal should not be to turn out products. Diverting public funds to develop unconnected collections of prototypes in the long-term undermines the science and skills base. We are lucky in the UK that our research councils still allow us to do generic, problem-oriented work.

FOR INFORMATION
Marc Moens is Manager of the Language Technology Group at the Human Communication Research Centre, University of Edinburgh. He can be contacted at:
Human Communication Research Centre
2 Buccleuch Place
Edinburgh EH8 9LW
Scotland, UK
Tel: + 44 131 650 4427
Fax: +44 131 659 4587
Email: M.Moens@cogsci.ed.ac.uk
WWW: http://www.ltg.ed.ac.uk

We'd like to hear what you think!

ELSNews exists to keep ELSNET members in touch with what's going on in the Language and Speech community, to inform them of new research and initiatives in the area, and to provide a platform for different views on topics of interest to our readership.

But *ELSNews* also wants to stimulate discussion and encourage the exchange of ideas. You may not always agree with the views put forward, or you may have something relevant to add. We welcome your replies, suggestions and comments! (Proposals for) feature-length articles are also welcomed, particularly if they fit in with the following themes:

- (June issue): **ELSNET in Wonderland: How can we turn ELSNET into a showcase of L&S technology?** *(deadline: 15 May)*
- (September issue): **Integration: Problem or Solution?** *(Deadline 15 August)*

Please send all correspondence to elsnews@let.ruu.nl, or to ELSNET (address at the back of this issue)

Future Events

Jun 2-5, 1997: UM97: Sixth International Conference on User Modeling, Chia Laguna, Sardinia, Italy. For information, contact um97-demos@cs.uni-sb.de. URL: http://www.cs.uni-sb.de/UM97

Jun 10-15, 1997: DIALOGUE'97, Conference on Computational Linguistics, Moscow. For information, contact dialog@artint.msk.su

Jun 16-27, 1997: LOT Summer School 1997, University of Amsterdam, Netherlands. For information, contact LOT@let.ruu.nl. URL: http://wwwots.let.ruu.nl/LOT/zs97.html

Jun 17-18, 1997: Speech and Language Technology (SALT) Club Workshop on EVALUATION IN SPEECH AND LANGUAGE TECHNOLOGY, Sheffield, United Kingdom. For information, contact R.Gaizauskas@dcs.shef.ac.uk. URL: http://www.dcs.shef.ac.uk

Jun 25-27, 1997: Third Workshop on Applications of Natural Language to Information Systems, Vancouver, British Columbia, Canada. For information, contact mcfet@cs.sfu.ca

Jul 2-3, 1997: Cognitive Science towards Applications (French Speaking), Villeneuve d'Ascq, France. For information, contact scicoia@univ-lille3.fr. URL: http://www.univ-lille3.fr/www/scicoia/

Jul 7-18, 1997: Computational Models of Speech Pattern Processing. Jersey, United Kingdom. For information, contact ponting@signal.dra.hmg.gb. URL: http://www.dra.hmg.gb/spp/sppasi.html

Jul 07-12, 1997: ACL/EACL 97: 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics. URL: http://horacio.ieec.uned.ed/cl97/

Jul 13-26, 1997: EUROLAN'97 Summer School in Corpus Linguistics , Iasi, Romania. URL: http://www.infoiasi.ro/eurolan97.html

Jul 14-25, 1997: ELSNET's 5th European Summer School on Language and Speech Communication, Katholieke Universiteit Leuven, Belgium. For information, contact ess97@ccl.kuleuven.ac.be. URL: http://www.ccl.kuleuven.ac.be/ess97/ess97.html

Jul 14-16, 1997: First International Workshop on Human-Computer Conversation, Bellagio, Italy. For information, contact DavidL@intrsrch.demon.co.uk. URL: http://www.dcs.shef.ac.uk/research/ilash/Meetings/Bellagio/

Aug 11-22, 1997: ESSLLI'97 (European Summer School in Logic, Language and Linguistics), Aix-en-Provence, France. For information, contact esslli97@lpl.univ-aix.fr. URL: http://www.lpl.univ-aix.fr/~esslli97

Sep 22-25, 1997: EUROSPEECH'97, Rhodes, Greece. For information, contact: tonesca@di.uoa.gr. URL: http://www.cti.gr/~ee-www/

ELSNET Participants

Academic Sites

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

Industrial Sites

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

What is ELSNET?

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

The long-term technological goal which unites the members of ELSNET is to build integrated multilingual NL and speech systems with unrestricted coverage of both spoken and written language. Building multilingual NL and speech systems requires a massive joint effort by two pairs of communities: on the one hand, the natural language and speech communities, and on the other, academia and industry. Both pairs of communities are traditionally separated by wide gaps. It is ELSNET's objective to provide a platform which bridges both gaps, and to ensure that all parties are provided with optimal conditions for fruitful collaboration. To achieve this, ELSNET has established an infrastructure for sharing knowledge, resources, problems, and solutions by offering (information) services and facilities, and by organising events which serve academia and industry in both the language and speech communities. In this respect, it is important to note that a network like ELSNET can only function well if all members of the network are prepared to give and to receive.

Electronic Mailing List

elsnet-list is ELSNET's electronic mailing list. Email sent to elsnet-list@let.ruu.nl is received by all Managing, Associate and Industrial node coordinators of the Network, as well as other person who are not necessarily members of ELSNET, but who have an interest in ELSNET's activities. This mailing list may be used to announce activities, post job openings, or discuss issues which are relevant to persons in the European natural language and speech communities. To request additions/deletions/changes of address in the mailing list, send mail to luz@cogsci.ed.ac.uk.

ELSNET's WWW Document

Detailed information about ELSNET and its activities and publications is available through the Web at the following URL:

`http://www.elsnet.org`

Comments and suggestions for new WWW pages are very welcome. In particular, each ELSNET site coordinator is encouraged to send details of his or her site's home page so that a hyperlink might be set up to it from the ELSNET home page.

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