# MT News International

*Newsletter of the International Association for Machine Translation*

**Published in February, June and October**

ISSN 0965-5476 Issue no.15, October 1996

**IN THIS ISSUE:**



***Editor-in-Chief:***
John Hutchins, The Library, University of East Anglia, Norwich NR4 7TJ, United Kingdom
Fax: +44 1603 259490. Email: J.Hutchins@uea.ac.uk; or: 100113.1257@compuserve.com
***Regional editors:***
**AMTA:** David Clements, Globalink Inc., 4375 Jutland Drive, Suite 110, San Diego, CA 92117, USA. Tel: +1 619 490-3680 x 302; Email: 71530.3476@compuserve.com
**EAMT:** Jörg Schütz, Institute for Applied Information Sciences (IAI), Martin-Luther-Str.14, D-66111 Saarbrücken, Germany. Fax: +49 681 397482;
Email: joerg@iai.uni-sb.de
**AAMT:** Professor Hirosato Nomura, Kyushu Institute of Technology, Iizuka, 820 Japan.
Fax: +81 948 29-7601; Email: nomura@dumbo.ai.kyutech.ac.jp
***Advertising Coordinator:***
Bill Fry, Association for Machine Translation in the Americas, 2101 Crystal Plaza Arcade, Suite 390, Arlington, VA 22202-4616, USA.
Tel: +1 703 998-5708; Fax: +1 (703) 998-5709.

*Published in the United States by Jane Zorrilla*

## From the Editorial Board

## Poll of MTNI Readers

In the last issue of MT News International we included a questionnaire inviting readers to give the editorial board their opinions of the content and direction of your newsmagazine. Thanks to all who have responded so far. In order to get a fully representative sampling of readership, we have decided to include the questionnaire again in this issue. Those who have not yet done so are urged to fill it in and send it to us as soon as possible. Otherwise you may find that we are leaving out items which you would like to have included, or we will not longer be reporting news which you would like to be reading about. It is your newsmagazine. Help us to meet the needs of all our readers.

So, please take a look at this issue and the immediately preceding ones, sharpen your pencil, and fill in your responses. The completed form may be faxed to the Editor-in-Chief or your regional editor at the address given on page 2. An electronic version may be requested from eamt@cst.ku.dk or AMTAInfo@aol.com.

# SPOTLIGHT ON THE NEWS

## Two MT systems at the Olympics

### Meteo Translates US weather reports

In January, the U.S. National Weather Service (NWS) handed the problem of the French translation of all weather reports during the Summer Games to the Chandioux Group, Montreal. On May 13, the Group delivered and successfully tested a special version of its machine translation system, METEO®.

English and French were the two official languages of the Games. The Georgia version of METEO ensured the accurate, rapid translation for all possible weather conditions using a lexicon of more than 3,000 weather-specific terms. While the special NWS version was new, the original system had already given ample proof of its superiority. Today, METEO translates some 27,000,000 words per year for the Canadian Government Translations Bureau, making the Chandioux Group its largest independent supplier of translation. In March 1997, the Translation Bureau will celebrate the 20th anniversary of its use of METEO for Environment Canada.

As John Chandioux, the Group's President said, "We are proud of the fact that, in the short time available to us, we were able to develop a translation product that provides 95% accuracy. It's like winning a first medal for Canada!"

### Globalink Joins Forces with Discovery Channel

Globalink, Inc. participated in a special Olympic-themed project with Discovery Channel Online (http://www.discovery.com). During June and July, Discovery Channel Online (DCOL) followed five Olympians from around the world as they prepared for, and participated in, the Atlanta Games. With Globalink's help, DCOL offered its multi-faceted online Olympic-themed feature story to both English and Spanish speakers around the world.

For this, the first of several projects with DCOL, Globalink dedicated a team of professionals from its 2,000-member translator network, and employed its language translation software, to translate DCOL's Olympic-themed special event into Spanish. The result was a bilingual site, allowing for wider global participation.

Said Thomas Hicks, publisher of Discovery Channel Online, "Our web site receives over 24,000 visits per day, many of which are from non-English speaking countries. As we move in the direction of making our site more globally appealing, this is a great first step. Our feature about five athletes competing in the 1996 Atlanta Games offers a special opportunity to launch the effort. Now we'll be able to reach the vast Spanish-speaking audiences around the world in their native language."

---

## Transparent Language Inc. Acquires Transcend from Intergraph Corporation

[Press release, July 1996]

Transparent Language, Inc., developer of the LanguageNOW! system of award-winning computer-based language learning programs, announced today that a newly formed subsidiary, *Transcend Language Corporation*, has purchased the Transcend Natural Language Translation technology from Intergraph Software Solutions, a division of Intergraph Corporation, located in Huntsville, AL. Transcend allows users to take text in natural language from a wide variety of electronic formats and automatically transform it into another language.

"Transcend is a formidable technology that not only provides us with an immediate presence in the machine translation marketplace, but also positions us to be a player in the global communications arena," said Michael Quinlan, President of Transparent Language. "We have always provided the best tools to learn a foreign language. Now, with Transcend, we will provide the best tools to instantly translate and comprehend a foreign language. This acquisition should position us to be the 'language company.'"

Gerald W. Pearson, Transcend's Director of International Sales and Business Development, is equally optimistic. "Transparent Language provides Transcend with the sales and marketing resources to bring this technology to a broad-based consumer market. We have already established international distribution channels and are currently negotiating a variety of OEM and bundling arrangements. The interest in the technology within the European market place has been overwhelming."

Released in March, 1995, Transcend is popular with professional translators. In addition, its corporate clients include McDonald's and AT&T, as well as CompuServe, which translates messages among over 50,000 users in the World Community Forum. Transcend version 1.2 requires a PC with an 80386 or higher microprocessor, the Windows 3.1, Windows 95, Windows for Workgroups, or Windows NT operating systems, 8 MB of memory, 15 MB of hard disk space for one language direction (20 MB for bi-directional installation). Microsoft Word 6.x or Word Perfect 6.x is required for integration.

Transcend Natural Language Translation version 1.1 is available for $495 for one language direction and $795 for a bi-directional kit (such as English/French and French/English). Language directions include: English-Spanish, Spanish- English, English-French, French-English, English-German, German-English, English-Italian, and English-Portuguese.

---

### Canadian Government Licenses Globalink Software

Globalink has licensed its bilingual translation management software to Statistics Canada, the entity responsible for collecting, analyzing and distributing Canadian census, economic and social statistics in both French and English. Globalink will help Statistics Canada manage more than 500 translation jobs each month with the use of Globalink's Translate Direct Management System (TDMS). Employees of Statistics Canada will be electronically linked by Globalink TDMS to the agency's internal translation service which will allow them to access any of its contracted translators.

Globalink TDMS is one of the company's newest offerings for corporate and government environments that deal in high volumes of professional translation jobs. Globalink TDMS allows the entire translation process to be centralized and monitored from a desktop PC. Employees ("clients") simply send jobs for translation via e-mail to Globalink TDMS, which automatically assigns incoming jobs to qualified translators who regularly query the system for work. Internal quality control officers are part of the automated process. At any time Globalink TDMS can report which jobs have been completed, which are in process and when they are due to be returned to the customers. The system ensures that jobs are completed on time and meet quality standards before returning them to the "client", electronically through the agency's e-mail system.

Larry Rousseau, Program Manager for Statistics Canada's Official Languages & Translation division said, "Globalink TDMS is already proving to be a highly efficient and cost-effective way to manage our immense translation demand. Last year our division handled over 7 million words; this year with Globalink TDMS we will be able to handle this volume, and more, with less administrative resources. Globalink TDMS simplifies time-consuming administrative and accounting procedures and allows us to produce all kinds of reports. This helps us to control costs and process a greater number of translation requests in a shorter period of time."

Globalink's vice president for Industry Marketing, Jeff Gray added, "Globalink is uniquely qualified to provide Globalink TDMS, because it is the same system we use to manage our own worldwide network of more than 2,000 translators. By customizing the system for Statistics Canada, we were able to gain a better understanding of the on-going translation issues facing the entire Canadian government. We are now prepared to provide translation solutions to other agencies and ministries, who are also required to conduct business in French and English."

---

# PRODUCTS and SYSTEMS

### Langenscheidts T1

[From publicity brochure]
Earlier this year, the publishing house of Langenscheidt in Munich, famous for its series of dictionaries, announced the appearance of a German-English and English-German translation system. Known as *T1*, the system is based on the Metal system developed by Siemens-Nixdorf for workstations and then further developed for personal computers by the Gesellschaft für Multilinguale Systeme (GMS). GMS itself was established originally for the development of a Russian-German system based on research at the German Academy of Sciences (of the former

GDR), and later for the development and research of already existing Metal systems in succession to Sietec (a subsidiary of Siemens Nixdorf).

The *Langenscheidts T1* is appearing in two versions: Standard (priced at DM.298), and Standard Plus (DM.398, announced for July 1996). Both versions come with the Metal system dictionary of some 320,000 entries with rich grammatical information. The difference between the two versions lies in the size of the supplementary dictionaries. In the Standard version, Langenscheidt provides its Taschenwörterbuch Englisch (120,000 entries); in Standard Plus, it provides the Handwörterbuch Englisch (220,000 entries). The publicity brochure stresses the ease with which texts can be translated sentence by sentence or en bloc directly out of Word 6.0 or 7.0. Unknown words are marked and alternatives presented. Also offered is a dictionary lookup facility for individual words. Several texts can be loaded and translated as a batch in succession. As in the original Metal system, *T1* provides a wide range of specialised subject dictionaries and facilities for user-specific extension of the 'system dictionary'. *Langenscheidts T1* requires an IBM-compatible 486 with 66MHz, or Pentium; 8MB RAM or more, 105MB minimum hard disk capacity, and a CD-Rom drive. It can run on Windows 95, Windows 3.1, Windows for Workgroups, and Windows NT 3.5 (with 16MB RAM).

---

## LogoVista E to J

### LEC announce upgrade and version for WWW translation

[Press releases, May 1996]

On May 31, 1996, Language Engineering Corporation announced upgrades to both the Personal and Pro versions of LogoVista™ E to J, the English to Japanese desktop system. Both are available for Microsoft Japanese Windows 95, Windows 3.1, Windows NT 3.51 or later, Macintosh and Power Macintosh. The new versions feature "much faster translation with improved accuracy, improved dictionary searching features, and many other new capabilities." LogoVista E to J Pro 3.0 also includes the capability of translating HTML and SGML files, and the storage of preferred translations of entire sentences in translation memory archives.

Language Engineering Corporation also announced *LogoVista E to J Internet***.** Working with the Netscape Navigator™ web browser this enables the translation of web pages from English into Japanese. On April 24, 1996, IBM signed a three-year agreement to bundle LogoVista E to J Internet with Aptiva personal computers sold in Japan. The product offers the choice of translating entire pages, selected text on a page, or only headings and hyperlinks. Users can select from among three speed settings for the desired combination of speed and accuracy. There are three display options: users can replace the English text with the translation, display the translation in a second browser window, or display each new translation in its own browser window. LogoVista E to J Internet is available for Microsoft Japanese Windows 95, Japanese Windows 3.1, Japanese Windows NT, Macintosh, and Power Macintosh.

For more information: Language Engineering Corporation, 385 Concord Avenue, Belmont, MA 02178 (Tel: +1 800 458-7267, +1 617 489-4000; Fax: +1 617 489-3850; Email: info@hq.lec.com; WWW: http://www.lec.com; AppleLink: LEC.)

---

## Systran for your Pocket

[Press release, August 1996]

SYSTRAN Software Inc. and SEIKO Instruments Inc. announce a technology transfer resulting in high performance hand-held translators. SYSTRAN and SEIKO have signed an agreement that will integrate SYSTRAN morphology, including linguistic data and software, into SEIKO's Translator Collection.

"This is the first time SYSTRAN's powerful technology will be available to the average consumer," said Larry Miller, general manager of SEIKO Instruments, USA, Inc. noting that the technology includes a massive dictionary of words and idiomatic expressions, verb conjugating ability, and inflection tables. "This relationship will result in products with greatly expanded capabilities, representing a new era in language translation technology."

Denis Gachot, president of SSI, said, "These pocket-size translators will be excellent travel companions for tourists, and they will meet the more demanding needs of business travelers, foreign language students and people struggling to function in the United States while they learn English. These small devices represent a breakthrough in cultural assimilation."

SEIKO translators with SYSTRAN technology will be introduced at the Consumer Electronics Show in Las Vegas in January. Shipments of the product will begin in March. The first language pairs offered will be English/Spanish and English/Portuguese, with others planned to follow. The SYSTRAN technology will be incorporated in a variety of models, priced from approximately $20 to $80.

Last year SYSTRAN brought its advanced mainframe technology to the PC with its SYSTRAN PROfessional program for Windows. SEIKO Instruments Inc. is a leading manufacturer of electronic hand-held consumer linguistic products, including translators, spell checkers and dictionaries; it dominates the English/Spanish translator market and has shipped more than 2 million hand-held units.

---

## Globalink Web Translator Shipped with Banyan's VINES

Banyan Systems Incorporated has chosen Globalink, Inc. to provide an exclusive software bundle for its VINES 7.0 Renewal Kit, shipping to VIP contract customers in October. By selecting Globalink Web Translator, Banyan is presenting its customers with the most complete Internet package available. Globalink Web Translator, the browser add-on to enable anyone surfing the Internet to translate sites currently shown in Spanish, French or German into English with the click of a button, includes AT&T WorldNet access software, 60 free hours annually of on-line time from AT&T's WorldNet service and Netscape Navigator 2.0.

The Globalink Web Translator package, with free AT&T WorldNet access and a special version of Netscape Navigator 2.0, will be shipped to almost 2,500,000 of Banyan's VIP customers by way of the VINES 7.0 Renewal Kit. This shipment will include a Globalink Web Translator licensing agreement for employees at each VIP site who wish to use the translation software. Globalink has extended special pricing of the Globalink Web Translator package to any VINES 7.0 recipient that wants the same Web translation capabilities at home. Corporate users--interested in connectivity and multilingual Web browsing at home--will be given a special ordering number and that will allow them to obtain this complete Internet offering for personal use, identical to the Globalink Web Translator package they use at work.

Glenn Pulling, Banyan's vice president of Business Development commented, "We've chosen to partner with Globalink because Globalink Web Translator with AT&T WorldNet Service enhances the value of the Banyan VINES for corporate customers needing to understand

multilingual sites on the World Wide Web. This is the first software package to be bundled with VINES and Banyan is pleased to be the first enterprise network provider to offer translation capabilities and Web access to its corporate customers."

Globalink Web Translator allows users to browse the Web as they normally would. Upon reaching a foreign language site, a user clicks Globalink Web Translator's "Translate" button, confirms the language pair (e.g., French to English), and after a brief pause the same page appears in the new language. Translated pages maintain all graphics, hotlinks and formatting. These translations can be created online, while surfing, or alternatively, pages can be downloaded and saved to be translated and viewed off-line.

---

## InterLan's EUROTRA: an interlingual PC-MT

*Jaro Lajovic*

InterLan/Geinsa, a company based in the Basque country, has recently completed the first (English-Spanish/Spanish-English) part of its multilingual MT system Eurotra. The system - not at all connected to the European project of the same name - deserves special attention as the first truly interlingual (soon to be marketed) commercial MT package. The French modules are under development; several other languages are to be successively added later. It is claimed that the development of both analysis and synthesis modules for any language would take from 0.5 to 1 linguist-year.

Eurotra has the general structure of an interlingual system: the source text is analysed by an inference machine using a source language database and rules, yielding language independent, disambiguated representation in proprietary MetaLingua. The target language database and rules enable the inference machine to synthesise the translation from the MetaLingua representation.

*Description*

The source program is written in C++. The existing system has a database of more than 150,000 concepts (claimed to correspond to a lexicon of approx. 500,000 words) per language. Translation can run either interactively or in batch mode. In the latter, Eurotra translates about 120 words per minute with Pentium 100 MHz and 8 MB of system memory.

There are two translation modes: automatic and semi-automatic. In automatic the general (all domains) dictionary is used; in semi-automatic mode, domain specific dictionaries (lexical sectors) can be selected. Currently, several domain dictionaries are available: computers, banking, finance etc. The user can also maintain his/her own user dictionary, while additional pretranslation memory (body of sentences/documents) allows translation of certain phrases, idioms, passages etc. in a pre-determined way.

The system can handle common file formats (such as ASCII, Microsoft Word, WordPerfect, HTML), preserving the format of the original document. The incorporated OCR enables automatic input of the scanned text. The ability to handle HTML makes Eurotra suitable for the on-line translation of WWW pages. Because of the conceptual analysis (needed to achieve MetaLingua representation) it is also said to be potentially useful as an intelligent information retrieval tool.

*General impression*

The author has been given the opportunity to test the system with the JEIDA test set (kindly provided by Dr. Hitoshi Isahara) and several medical and computer texts. All were

translated in automatic mode (ie. using the general dictionary), and one text was back-translated (English → MetaLingua → English). Back-translation - an interesting and useful feature - is available in the developers' version of Eurotra.

Evaluation from the user's point of view included the translation of five fairly demanding texts: two general medical (patient information), two specialised medical (research reports), and a computer one (part of a software manual). The first two translations were rather good (without using a specialised lexical sector), while the translations of the research reports were partly acceptable, as the abundance of specialist terminology hindered the processing. The computer text translation was quite good; its English source was also back-translated. Interestingly, some phrases in the English target text were rendered a little differently, but correctly. Back-translation seemed especially useful in pinpointing causes of inadequacies in translation.

In the evaluation made from the developer's point of view (with the non-standardised JEIDA test, ie. without model translation), Eurotra scored higher in the structural analysis part than in the disambiguation part. In the former, it scored highest in the parts of speech subgroup. In the latter, a discrepancy was noted: while disambiguation of some sentences failed, it succeeded in others of the same subgroup. In all the groups, the faults seemed partly to be caused by the system's attempt to relate consecutive sentences, which were not related in the JEIDA set.

Eurotra seems to be a most interesting system, especially because of its potential multilinguality. Besides several strong points (eg. contextual analysis, a sort of "soft fail" mechanism, word order) there are a few which are yet to be improved (eg. the correction of certain idiosyncrasies, perhaps the addition of certain heuristics). In general, however, the noted difficulties seemed minor. Because of the relatively easy creation of additional language modules, the system seems rather attractive - both for use with languages which are already MT-covered as well as with minor, not yet covered ones.

*System requirements*

Eurotra is to be delivered on CD ROM and installed on PC with 486 or higher (Pentium recommended), running under DOS 6.0 or later, Windows 3.1/95/NT or OS2, with min. 8 MB system memory (16 MB recommended; 512 kB cache recommended) and 130 MB (per language pair) on hard disk.

For further information, contact: InterLan/Geinsa, C. Legazpi 6, 48950 Erandio-Bilbao (Bizkaia), Spain (Tel: +34 4 467 6091; Fax: +34 4 467 6323; E-mail: geinsaeu@sarenet.es).

---

## Machine Translation Service Unit (MTSU) Provides Networked MT from Singapore

Based at the Institute of Systems Science, National University of Singapore, MTSU is providing a high-speed service for large-scale translation from English into Chinese, Malay and Japanese. The service was launched at the MT Summit V in Luxembourg in June 1995, and attracted much interest at the Forum of LISA (Localisation Industry Standards Association) in Singapore during October 1995. Its speciality is the localisation of software and its documentation and already includes among its clients Informix Asia/Pacific, Fisher Rosemount, Holistic Systems, and Dataware Technologies. The MT system developed at ISI over six years forms the backbone of the service, supported by professional translators to achieve high quality output. A particular strength is translation into Chinese: available as either Simplified Chinese (for China and Singapore) and Traditional Chinese (for Taiwan and Hong Kong). Shortly MTSU will offer also

translation into Korean, and translation from Japanese into English. Translations can be delivered as electronic files, hardcopies, postscript files, films of Chinese DTP files. For software messages, the translations are delivered formatted and ready for immediate software integration. Translation of web pages is now available in partnership with Star+Globe Technologies.

For further information contact: Ms Jane Ee, Manager, Machine Translation Service Unit, Institute of Systems Science, National University of Singapore, Heng Mui Keng Terrace, Kent Ridge, Singapore 0511 (Tel: +65 772 6696; Fax: +65 774 4998; Email: jane@iss.nus.sg)

# CONFERENCE REPORTS

## EAMT Machine Translation Workshop
### 29-30 August 1996, Vienna

*John Hutchins*

The workshop was organised by the European Association for Machine Translation in conjunction with the fourth international conference on Terminology and Knowledge Engineering (TKE'96) and took place at the Technical University in Vienna. Held over one and half days, the theme of the workshop was "Machine translation: language resources, terminology, economics and user needs" and successfully attracted over 50 participants, many from organisations already involved in applications of machine translation and other computer-based translation tools and workstations.

After introductions by John Hutchins (president of EAMT) and by Dimitri Theologitis (Translation Service, CEC), the organiser of the workshop, the first talk was given by Reinhard Schäler (Localisation Resources Centre, Dublin, Ireland). He began by describing the particular needs of the localisation industry, which has a considerable presence in Ireland (the largest concentration in Europe, and perhaps the world), and outlined the background to the establishment of the Localisation Resources Centre. For localisation translation, he stressed the particular importance of translation memory (TM) facilities: all efforts devoted to the post-editing of MT output were effectively lost unless the resulting translations are stored for future use and consultation. However, although 70% of texts are found to be unchanged, and a further 15% are sufficiently similar to be located by a TM facility, there cannot be complete reliance on this method. There was a need for storage and alignment of phrase segments (a 'phrasal lexicon') in conjunction with MT-type analysis and generation; and research on these lines is in process for German and English at the University College Dublin. He was followed by Michiel de Koning (Cap Volmac, Utrecht, Netherlands), whose talk was devoted primarily to the work on controlled languages and the translation services provided by Cap Volmac.

In the second session Svetlana Sokolova (AO PROject MT Ltd, St Petersburg, Russia) reported on the results of a questionnaire sent to users of her company's *Stylus* software for translation from English, French and German into Russian, and vice versa. The product has over 15,000 users including many large corporations in the United States and Europe. The majority (86%) of users were post-editing output; a large number (46%) were using the system to browse texts for information assimilation; and some (6%) were using Stylus as a check on the correctness of a human translation (e.g. a letter in English written by a Russian). Most translation was of technical material (70%), but contracts (40%), correspondence (65%), scientific articles (44%), and software documentation (32%) were well represented.

Annelise Bech of Lingtech (Copenhagen, Denmark) described the use of the *PaTrans* system for translating English patents into Danish. Lingtech is a translation company set up by the Danish patent agency Hofman-Bang & Boutard, Lehmann & Ree A/S, specialising in the translation of patents from English, German and French into Danish. *PaTrans* (developed by the Center for Sprogteknologi (CST) on the foundations of experience with Eurotra) is used for 75-80% of all the translations from English into Danish. Decisions about suitability for MT are made on the basis of complexity and newness of terminology, and the quality of writing (some patents are written by non-native speakers). Lingtech maintains special dictionaries for the system (the general dictionary was set up by CST); initially it was expected that there would be many special dictionaries, but experience has shown that in practice only two large ones, for chemical and 'mechanical' terminology, and three small supplementary ones are required. It is difficult to classify terms into subject fields and to prioritise dictionaries for many texts, and as a consequence, dictionaries are more heterogeneous than anticipated.

After lunch came a session devoted to the experiences of the European Commission's translation service. Jean-Marie Leick (DG XIII, Luxembourg) concentrated on the integration and maximised utilization of the rich lexical resources within the Commission, and in particular the work on the creation of the *Euramis* Linguistic Resources Database. He described the technical and linguistic problems of importing Eurodicautom (600,000 multilingual entries) into Systran, and the exportation of Systran dictionaries into the Euramis database. Dorothy Senez (Service de Traduction) then reported on recent activities on behalf of the Translation Service of the Commission: an in-house user survey, a market study (including experiences of users), practical experiments (to demonstrate that MT is an asset to the Translation Service), legal issues (concerning Systran and the Commission), and a cost-benefit study (to take place in the coming autumn). A full account is given elsewhere in this issue of MTNI.

The first day was concluded by a description from Harri Arnola of the MT technology of Kielikone Ltd. and an outline of the *TranSmart* system for Finnish to English translation, which combines TM and MT. After arguing that translation systems are necessarily incomplete and never exploit even their own capabilities to the full, Arnola proposed a method of evaluating performance based on the percentage of correct sentences as a proportion of the total theoretically achievable by a system. Then Martha Ebermann gave an illuminating account of how Coop Schweiz introduced and applied the *Trados* workstation for translating German documents (mainly recipes, packaging texts, and manuals) into French and Italian. Such has been the volume of material translated that use at Coop Schweiz is taxing the memory capacity of Trados to its limits. She estimated that outlay for the system introduced in August 1993 had been recovered by the end of 1995. Finally, Milde Jordaan-Weiss (of the South African National Terminology Service, Pretoria) gave a brief description of the MT systems marketed by Epi-Use, for translation from English to Afrikaans, German and French, and vice versa, from Portuguese and Italian into English, and the development of systems for translation involving South African languages: Zulu, Xhosa, Swahili, Tswana, Sesotho and Sepedi. The Afrikaans, Portuguese and other European language systems are sold at 1,500 Rand in South Africa and scaled-down versions will be marketed at $100 in Europe.

The second day began with Lee Humphreys (GSI-Erli, France) giving details of the translation tools developed and under investigation at *ERLI*: AlethTR for French-English and English-French translation, integrating TM and a traditional transfer based MT system; the development of lexical tools on the basis of the theory-independent 'generic lexicon' database GENELEX; AlethKES, a workbench under development for automatic knowledge and

terminology extraction from corpora; and finally the outlines of a new translation engine (L6) exploiting experience of GRAAL (a re-usable grammar formalism). He was followed by Matthias Heyn (Trados Benelux) arguing that the TM approach (as in the *Trados* workbench) is far more appropriate in the professional translator environment than any conceivable MT system, because it starts from an expansion and elaboration of traditional work patterns, it exploits the results of previous human translation, and it provides access to problem areas of terminology. MT is not excluded as an additional aid, as long as it is incorporated a mechanism for 'proposing' translations and not for presenting 'definitive' translations which have to be edited.

A detailed account of an evaluation of MT and translation tools on behalf of the Black and Decker company was presented by Adriane Rinsche (Language Technology Centre Ltd., UK). Valuable insights were given of working practices and assumptions about translation, which may be typical of other multinational companies. The final recommendation was for translator's workbenches, but the company decided to take up an outsourcing option and translations are now undertaken on its behalf by Rinsche's own organisation.

The final session of the workshop was devoted to the impact of networking on MT. Joachim Meyer (Logos GmbH, Germany) described the internet service that Logos will be inaugurating shortly; and Jörg Schütz (IAI, Saarbrücken) concluded the workshop with a survey of the future 'networked computing' and its potential implications for MT and translation services. Current developments in company intranets and MT client/server systems indicate future directions. At the core will be cooperation among MT providers and translation brokers to identify specific user needs, to ascertain levels of quality and affordability, and to evaluate benefits and value of translations and MT services. The message for MT and translation technology providers was clear: the demand for on-line immediate translation services will grow enormously, and they must be ready to meet it.

The full proceedings of the workshop will be published by Dimitri Theologitis on behalf of EAMT; details will be announced in the next issue of MTNI.

---

## Terminology and Knowledge Engineering (TKE'96)

### 4th International Congress, 26-28 August 1996
### Vienna, Austria

*John Hutchins*

The TKE 96 conference attracted some 170 participants from all over the world to its venue at the Technische Universität in Vienna. Organised by InfoTerm under the chairmanship of Christian Galinski this was the fourth TKE conference on a topic which has obvious interest and relevance to the MT field. Sessions were devoted to Terminology and Language Engineering, Terminology and the Philosophy of Science, Terminology and Knowledge Data Management, Terminology on the Information Superhighway, and Terminology and Translation. Most talks had some relevance to MT, given the importance of compiling and maintaining lexica conforming to international terminological standards and the common interests in building bilingual and multilingual text corpora. The full proceedings were made available to participants at the beginning of the conference in a book published by Indeks Verlag [See Publications Received.] Of direct relevance to MT the following papers should be mentioned.

Walther von Hahn reported on developments of the knowledge-based MAT system for German and Bulgarian funded by Volkswagen and being developed at Hamburg with researchers from the Bulgarian Academy of Sciences. A paper by Jörg Schütz (IAI, Saarbrücken) entitled 'Terminology and information engineering' recounted experience of a project at BMW (Munich)

on developing tools for multilingual access to a technical information service for the support of car maintenance. Eva Dauphin (Aérospatiale) reported on the TRANSTERM project, which ended in June 1996, for a terminology toolbox based on a linguistic model which supports the integration of terminological data and lexical data conforming to the GENELEX model. The toolbox will support applications in controlled languages, automatic indexing and machine-aided translation (e.g. as part of the Aleth project at GSI-Erli).

Masumi Narita and others from Ricoh (Japan) described the construction of *Tanyakuman*, a multifunctioning 'language assistant' comprising an optical character reader, a language processor, and a layout controller. The language processor consists of an English-Japanese dictionary, a lookup engine, and a part-of-speech analyser. From an English document input, the system produces a photocopy including interlinear Japanese translations of individual English words (mainly nouns). In most cases, two alternatives are offered. As a result, Japanese readers are able to work out the basic content of the English texts. The lexical database is derived from Ricoh's resources used in the development of an MT system at Ricoh. There are limitations on the size of the dictionary imposed by memory constraints, but nevertheless it contains 57,000 English words with their most common Japanese equivalents. An evaluation of the word translations indicates significant improvements by Japanese in reading comprehension of English texts.

The conference began and ended with keynote speeches (not included in the proceedings) by Khurshid Ahmad -- replacing Yorick Wilks --devoted to problems of tagging corpora of sense discrimination; and by John Sowa on the development of ontologies relevant to terminology and knowledge-based language engineering.

## NLP conference at Moncton University

**4-6 June 1996**

In the first week of June, the Université de Moncton (New Brunswick, Canada) hosted an international conference on "Natural Language Processing and Industrial Applications" organised by Chadia Moghrabi and GRÉTAL, Groupe d'étude sur le traitement automatique des langues of the University.

A number of the contributions were devoted to MT-related topics, including the opening invited presentation by Jaime Carbonell ('Software engineering approach to machine translation') and the second invited talk by Christian Boitet ('La synérgie entre THAM, réseau et TA comme facteur de progrès théoriques et pratiques en TAO'). For example, Manny Rayner and David Carter spoke on 'Adapting the Core Language Engine to French and Spanish', Rémi Zajac on 'A multilingual translator's workstation for information access', Katty Grasson on 'Une maquette d'environnement de typage textuel pour la TA fondée sur le dialogue' (on the LIDIA project). Other contributions involving translation included E.Agirre and others on 'Design of a translator-oriented dictionary', Chadia Moghrabi on 'Portability in a text generation system', Kurt Godden on 'Statistical control charts in natural language processing', and Jessie Pinkham on 'Grammar sharing between English and French'. Finally there were three contributions covering evaluation issues: 'Building bridges for translation tools' by Reinhard Schäler, 'Test suites for quality evaluation of NLP products' by Frederik Fouvry, Lorna Balkan and Doug Arnold; and 'Working towards user-oriented evaluation' by Sandra Manzi, Maghi King and Shona Douglas.

# USERS and RESEARCH

## News from the European Commission

### Machine Translation User Survey

*Dorothy Senez*

The Translation Service of the European Commission is currently examining the conditions under which machine translation might continue to be funded within the institution in the coming years. To this end, it has launched a feasibility study encompassing the following separate, but interrelated, points:

(1) an in-house survey of Commission MT users, to ascertain their MT needs in regard to languages, speed and quality;
(2) practical experiments with the help of in-house translators to provide objective data on the effects of machine translation on the Translation Service's production line;
(3) an investigation into the legal issues which dictate the use of the Systran system by third parties;
(4) a technological survey of the state of the market for MT products and services, the aim being to establish whether there are alternatives on the market to the Commission's current MT system;
(5) a cost-benefit analysis based on the results of these first four studies.

This article will focus on the survey of in-house users of MT. Other aspects of the feasibility study will be discussed in forthcoming issues of MTNI.

Faced with the decision to continue, or not to continue, with MT at the Commission, it was felt that important strategic guidance could be provided by the experiences of users. There were two groups to be considered:

- in-house Commission translators;
- administrators in the other operational departments of the Commission.

While the survey was an ideal opportunity for MT users to voice their opinion on the service offered, the remarks of those who had never used MT would also provide valuable insights. Consequently, both users and non-users in each of the two groups were consulted.

*The Questions*

With the user survey came the opportunity to establish some objective data on machine translation, thereby dispelling some of the myths surrounding the MT question. MT is freely available to all Commission officials via the internal electronic mail network and the majority of users are non-linguist staff who help themselves to machine translation as and when they need it. Consequently, there is no direct feedback. We set out to discover why, how, and how much MT is used, and who the users are. We also wanted to find out how useful it is to the Translation Service directly (texts post-edited by translators) and indirectly (texts submitted for MT by administrators that would otherwise have found their way to the Translation Service). It was also important to evaluate how the Commission departments rate MT as a linguistic tool (for browsing, translating, drafting). The survey would highlight its strengths and weaknesses and indicate the reasons why many people do not use MT. Should it prove to be a tool worth

maintaining, ways of improving the service would then need to be explored.

*Response Rate*

Questionnaires were sent to all 1,700 members of staff in the Translation Service, 520 of whom had been identified as MT users. In all, 773 responses were received. A total of 2,600 users in the administrative departments were surveyed, with 735 responses received. In the case of non-users a proportional sampling method covering 5.5% of the remaining Commission staff was adopted and 270 responses were received.

*Preliminary Results*

At the time of writing only preliminary results can be presented. The final conclusions, however, will enable decisions to be taken regarding the future orientation of MT at the European Commission.

*Translation Service - Users*

Translators like to use MT because of its speed, the typing time it saves,and the fact that the raw MT is returned with its original format. MT is also sometimes used to provide assistance with terminology. Although translators are generally somewhat negative about the quality of the MT output, a certain number did appreciate the system's sense of humour. As nearly all those translators who use MT do so to produce a final, polished, translation, they do not like the heavy post-editing involved. The majority find less than half the documents useful, and a small percentage find no texts useful. Over 50% of users in the Translation Service request MT only occasionally, and although they do save some time, a significant number said they saved very little time. Everyone at least agrees that the system's response time is very good. As regards the assessment of language pairs, the best marks are attributed to French-English, French-Spanish, French-Italian, and English-French. About half the users in the Translation Service would like to able to create and manage their own personal dictionaries. Although only 25% of translators said they would not have met deadlines if MT had not been available, the majority (67%) felt MT was a tool worth having at their disposal.

*Translation Service - Non-users*

Nearly half of non-users say they do not know how to use the system. Many feel that their texts are not suitable, and others fear a dehumanising effect on their work. In many cases the relevant language pair is not covered.

*Administrative departments - Users*

Administrators tend to request MT on an occasional basis for the translation of urgent documents they would have preferred to send to the Translation Service. They also use MT is for information scanning in languages unknown to the reader and for drafting in a foreign language. They like MT for its speed, ease of use, the lack of bureaucratic procedures and the fact that it is available round the clock. They do not, however, like having to correct the texts, the fact that the system is slow to learn, and that some language pairs are missing. Texts are revised in most cases, normally by a native speaker. More than half of the respondents in this group do not indicate on text that it is revised MT output. Those who rely on the post-editing service are happy with it. On the assessment of language pairs, administrators tend to be more lenient than their translator colleagues. The majority find half or more of the output useful. In stark contrast

to the Translation Service figures, 74% consider MT saves them a considerable amount of time. More decisively, over 50% think that some of their documents would have been late had it not been for MT. Also interesting, since it proves that MT is at least saving the Translation Service some time indirectly, 59% say they would otherwise have sent their texts to the Translation Service for translation. A resounding 94.8% feel that MT is worth having at their disposal.

*Administrative departments - non-users*

About 20% of non-users have no real need for translation. Of the others, most do not know how to use MT and those who do complain about poor quality, determined from direct experience in the past or from the comments of colleagues. As with non-users in the Translation Service, many feel that their texts would not be suitable. In some cases people solve their day-to-day translation problems with the help of colleagues. The majority of non-users requested more information about MT, which is still relatively unknown.

There are two quite different pictures emerging here. On the one hand there is a lukewarm, but by no means entirely negative, reaction from professional translators. Although results vary from one language group to another, a significant number have been able to make MT work for them, particularly in the case of targeted development, where the Systran dictionaries have been programmed for specific text types. On the other hand, there is a more enthusiastic reaction from the administrative departments. This group has perceived needs for urgent translations, browsing and drafting, which MT is already meeting to some extent. Nevertheless, the initial findings show quite clearly that increased efforts are required here to provide better support and information for this user population.

---

## European Research and Development in Machine Translation

*Jörg Schütz*

*1. Introduction*

In MTNI #14 Paul Hearn contributed a column with an overview of "Current Applications in Europe". We got some insights about current machine translation (MT) applications and MT trends such as Web-based translation environments as intended to be offered by the European OTELO project, as well as recent developments in the field of Controlled Languages (CLs) within the SECC project. In industry CLs are seen as being a prerequisite of large-scale good machine translation throughput. Paul has already indicated that the field of machine translation is diversifying.

Today, we can no longer talk about machine translation projects in isolation, but about projects making use of multilingual language technology which as a concept subsumes MT. In recent European projects this concept plays an important role and can be found in nearly every project description across different domains and application fields. Obviously, Europe is the biggest market for multilingual language technology products such as multilingual document management systems, multilingual information retrieval systems, etc., but there are also new markets emerging in the Eastern and Asian countries. Currently the languages of these new markets are only represented in a few European projects such as ARAMED which is concerned with the translation of medical diagnosis in Arabic, English and German, and some INTAS and Copernicus projects and concerted actions such as TELRI (Trans-European Language Resources Infrastructure). ARAMED is based on some of the results of the recently finished European

project ANTHEM (cf. MTNI #11, June 1995).

Regarding multilinguality we also have to mention the ubiquitous information networks, particularly the Internet and its multimedial extension, the World Wide Web (WWW or Web), that are claimed to be the driving force of the information society. All Web-centric applications are more or less faced with the problem of overcoming the existing language and cultural barriers to allow for effective information communication and exchange. This then is one of the foremost aspects of the ongoing European projects which I will introduce briefly in the following sections.

Most of these projects are concerned with information retrieval and information filtering with a focus on categorisation and knowledge extraction. Only some projects such as the German VERBMOBIL project and the European projects APOLLO, MAITS and OTELO are concerned with real machine translation (MT) capabilities, i.e. they make an explicit reference to translation.

*2. MT-centric projects*

The aim of VERBMOBIL is the speech-to-speech translation of face-to-face dialogues in the domain of appointment scheduling. It is now completing its first phase (1993-1996) with the Verbmobil research prototype FP 1.0. VERBMOBIL consists of a modular system architecture which allows three kinds of processing of the recognised speech input: (1) a shallow linguistic processing based on identified speech acts, (2) a deep linguistic processing, and (3) a combination of (1) and (2). Currently, the English language is used as an intermediate language, i.e. the dialogue language of the users of the VERBMOBIL device; in the second phase of VERBMOBIL (1997-2000) this will be replaced by a direct translation between the languages involved (German, English and Japanese). To date VERBMOBIL's vocabulary comprises approx. 2,500 words which will be extended to 10,000 words in the second phase. During the development cycles the project has put a strong emphasis on the control of language coverage, system throughput and translation quality; the latter with special attention to end-to-end evaluation cycles (black box validation and assessment). In its second phase the project envisages extending the application scenario to a multimedial telecooperation system (e.g. for distance working cooperation) and to dialogue situations with more than two participants (multimedia and multiparty capabilities). VERBMOBIL is coordinated by the German Research Center for Artificial Intelligence (DFKI GmbH) in Saarbrücken, Germany. The consortium consists of 29 partners from industry and academia.

APOLLO (An Open Workbench for Multilingual Document Creation and Maintenance) aims to provide an integrated documentation and document production environment in the domain of training materials for banks. Initially the system will handle only French and English but it will be designed to enable the incorporation of other languages at a later stage. The machine translation engine employed in the project is the CAT2 system which was developed as a Eurotra sideline at IAI Saarbrücken. The APOLLO consortium is led by Office Future International Services S.A.in Luxembourg.

The goal of the MAITS (Multilingual Application Interface for Telematic Services) project is to develop an Applications Program Interface (API) to support access to software based translation services. These services are intended to include codeset conversion, transliteration and cultural string conversion, translation memory, and machine translation. The API is intended for use within telematics service applications, such as messaging, to permit online multilingual capabilities. The institution responsible for the MAITS project is Sybase

France located in Paris.

*3. MT related projects*

The MT-related projects with a European dimension are mostly funded by the European Union under the Language Engineering and Telematics Applications Programme of the European Commission within the Fourth Framework Programme. In this programme we distinguish four clusters:

(1) Language Engineering Resources
(2) International Business Support
(3) Information Services
(4) Public Interest Support

In the following, I will describe some of the ongoing projects within these clusters. The projects belonging to the first cluster are all funded under the Second Language Engineering actionline (LE2) of the European Commission. The remaining clusters are funded under LE1. Currently, the Commission is negotiating with organizations who have submitted project proposals for LE3. We will report about these projects within this column of MTNI when appropriate information is available.

*3.1 Cluster 1: LE Resources*

The building of language resources can benefit different language technology applications including machine-aided translation and machine translation. Therefore this cluster is also of interest for the MT community, particularly since the projects of this cluster aim to provide the basis for standardisation efforts.

The EUROWORDNET (Building a Multilingual Wordnet Database with Semantic Relations between Words) project aims to produce a multilingual database with several links between entries based on relations such as antonymy, homonymy, entailment, cause, etc. These links can be monolingual and across languages. The database is intended to enrich and enlarge the existing Princeton WordNet for American English. Such a multilingual database can be used in a variety of applications, including machine-aided translation and quality information retrieval. The Computer Centrum Letteren of the University of Amsterdam is responsible for the implementation of the project.

INTERVAL (Interlinguistic Terminology Validation) will produce methods and tools for the validation and the standardisation of terminological resources. These resources are essential to a variety of applications, such as translation, document management, and software localisation. INTERVAL intends to reduce the costs of terminology creation and maintenance, to facilitate reusability, to ensure compatibility between different sources and to establish the basis for resources comprising high quality. CL Servicios Linguisticos in Madrid is the leader of the INTERVAL consortium.

LE-PAROLE (Language Engineering - Preparatory Action for Linguistic Resources Organisation) will produce a large-scale harmonised set of corpora (20,000,000 words) and lexica (20,000 entries per language) for all European Union languages. The resources will be produced in a standard format supporting selection and customisation (there seems to be an agreement on the GENELEX format). They will have a wide range of applications: for IT developers they will serve as extensive basic resources, which would be expensive to produce, and which can be used in developing and testing applications; for publishers, they will support the production of language learning material; and for academic research they provide both basic

resources and a basis for comparative analysis. The Consorzio Pisa Ricerche (CPR-CNR) in Italy is responsible for the language engineering part of the project and GSI-Erli Paris will provide the appropriate tools for the implementation of the project.

SPEECHDAT (Speech Databases for Creation of Voice Driven Teleservices) will produce speech databases with an extensive coverage of both languages and application environments. The databases will cover all 11 European languages and Norwegian, Slovenian, Welsh, and specific variants of Dutch, French, German and Swedish. They will include 5,000 speakers per language, covering various applications, speaking styles, and environmental influences. The Siemens AG in Munich coordinates the project.

The last project in this cluster aims to institutionalise the above mentioned efforts for the creation of resources. ELRA, the European Language Resources Association, aims to provide the needed infrastructure for identifying, collecting, classifying, validating, distributing, and exploiting language resources. These resources include both speech and text, basic data (corpora, recordings, terminology), linguistic models (grammars, lexica, speech) and software tools. Additional activities include the development of evaluation guidelines, central clearing house activities and the brokering between producers and users of resources. The ELRA office is located in Paris, France.

*3.2 Cluster 2: International Business Support*

Within this cluster we have the projects APOLLO, MAITS and OTELO which were already described above, and MABLe.

The MABLe (Multilingual Authoring of Business Letters) project is developing a facility to support the production of business letters in a foreign language which is known only to a limited extent by the author. The tool will guide, inform, interrogate, and correct its users so that the letters they produce are of high quality, taking into account not only the purely linguistic aspects of the communication but also variations in commercial practice. The final system will be multilingual and able to learn from its own use. This project provides a kind of machine-aided translation functionality. The responsible institution is Mari Computer Systems Ltd. in Gateshead, England.

*3.3 Cluster 3: Information Services*

This cluster is the biggest cluster in terms of the number of projects, which obviously reflects the key themes of the 1994-1998 work programme of the Commission. The topics covered by the following projects are:

(a) Human-computer interaction which includes spoken and written language based access to information and transaction systems.
(b) Person-to-person communication which subsumes foreign language learning, computer-aided translation, multilingual messaging and conferencing.
(c) Information retrieval and filtering for the provision of customised and 'profiled' information.

The ACCESS (Automated Call Center Through Speech Understanding System) project is designing and developing a call-centre which is able to deal with incoming telephone calls automatically by recognising and understanding what the caller is saying. The call centre will be used for handling particular situations such as dealing with callers who are making enquiries as a result of a specific advertising campaign. In this way the scope of the vocabulary used and the dialogue to be managed is not too broad and the call can be directed effectively. This project is

coordinated by the Daimler-Benz AG in Ulm, Germany.

CAVE (Caller Verification In Banking And Telecommunications) is developing a facility for recognising a speaker for the purposes of authenticating his identity. The results of the project will provide protection for services which are offered over the telephone and which require a high level of security and protection against unauthorised access. A typical type of service that would benefit from CAVE is telephone banking. The PTT Telecom BV located in Den Haag, the Netherlands, is the responsible institution.

ECRAN (Extraction of Content: Research At Near-market) will analyse free texts (initially, financial information from specialised newswire services, and market information on the Internet). A given text is compared against a predefined model of user requirements (user profile) in order to precisely identify relevant information on behalf of a customer. ECRAN includes 'lexicon tuning' which enables the system to be automatically extended and adapted to other domains. The ECRAN consortium is led by Thomson-CSF in Orsay, France.

EDITO (Digital Handling and Distribution of Personalized Press Clippings) is developing a service which will provide online press cuttings according to specific customer requirements. The customer has to specify the subject matter of interest. The service then applies this to a multisource database containing all newspapers which are collected electronically at the time of publication. Using semantic modelling techniques and full text indexation techniques, the service will be able to provide accurately profiled information. Axime Services of Paris, France, is responsible for the implementation of this project.

FACILE (Fast and Accurate Categorisation of Information by Language Engineering) will handle message dispatching and routing of texts for financial institutions (banks, trading companies, rating companies). The system will integrate shallow and deep text analysis in a robust and efficient tool, supported by a rich semantic domain model and a powerful text preprocessor. This kind of categorisation can be compared to that obtainable by a human quickly skimming through the text, recognising main topics dealt with but without actually reading the text. The Milan based software house Quinary SpA in Italy is the leader of the FACILE consortium.

MAY (Multilingual Access to Yellow Pages) will provide multilingual access to yellow pages on the World Wide Web. Customers enquire in their own languages and are presented with appropriate entries from the yellow pages. The system uses the meaning of the request for searching, so that the user does not have to be familiar with the language, terminology or traditional headings and indexes used in yellow pages. The service will be useful for searching both national and international yellow pages. The language resources employed in this project are based on the GENELEX format. GSI-Erli acts as tool provider and coordinator of the project.

MULTIMETEO (Multilingual Production of Weather Forecasts) is developing a system for generating weather forecast information in a number of different languages, from basic meteorological data provided by the weather forecasting computers of Meteorological Institutions. The objective is to deliver weather forecasts to customers in a way which is tailored to their particular needs in the language of their choice. Such a service is valuable to transport services, travellers and sports organisers for example. Again, GSI-Erli coordinates this project.

RECALL (Repairing Errors in Computer Aided Language Learning) is investigating the requirement to provide error correction as an enhancement to a commercially available Computer Aided Language Learning package. The facility will detect errors made by students and provide them with feedback in a way which helps them to understand the nature of the mistake which has been made and to learn not to make it. Initially the facility would be available

to learners of the English and German language. The Institute for Logic and Linguistic of IBM Germany in Heidelberg is the leader of the RECALL consortium.

REWARD (REal World Applications of Robust Dialogue) is developing a workbench for the development and implementation of dialogues to be used in an automatic call centre. The purpose of the workbench is to provide a tool which is suitable for use by people who are responsible for the function which the call centre is handling so that they no longer need the services of a technical expert to program the call centre for each application. Vocalis Ltd. Cambridge, England, coordinates the project.

SPARKLE (Shallow PARsing and Knowledge extraction for Language Engineering) will develop tools for phrasal-level syntactic analysis of naturally occurring text, which can be easily parameterised by language, and a system for semi-automatic lexical acquisition. The first of these will give accurate and immediate access to information in a variety of languages. The second will allow cost effective creation of lexicons which are sufficiently rich to be used in a range of other Language Engineering applications, such as multilingual information retrieval. CPR-CNR Pisa, Italy, is responsible for the implementation of this project.

SPEEDATA (Speech recognition for data-entry applications) is developing a system for data entry to databases containing structured and textual data, using continuous speech as the main input medium. The user interface will offer a robust system able to adapt to speaker variation and dialect and to operate with a range of languages (initially Italian and German). The initial application is to enable the rapid computerisation of a bi-lingual, Italian/German, land registry in North West Italy. The Italian software house Informatica Trentina SpA in Trento is the SPEEDATA consortium leader.

VODIS (Advanced Speech Technologies for Voice Operated Driver Information Systems), coordinated by the German Robert Bosch GmbH in Stuttgart, is developing a means of directing in-car devices using the human voice. Based on an existing in-car entertainment centre the project will enable the driver to control the functions of the centre, a GSM telephone, as well as an information system for working out the best routes for a journey. Eventually, it is expected to include information related to the vehicle about traffic congestion so that new routes can be selected.

### *3.4 Cluster 4: Public Interest Support*

The AVENTINUS (Advanced Information System for Multinational Drug Enforcement) project will define the requirements and propose a design of an information system to support the collection and analysis of intelligence useful in the prevention and detection of drugs related offences. Information will be available from a wide variety of sources, in both structured and textual form. AVENTINUS will investigate how such information can be collated, profiled and presented to investigators in a way which can effectively support their enquiries. GMS GmbH Munich, Germany, is the responsible coordinator of the project. It is intended that EuroPol (European police) authorities will join the project in the near future.

HAGIS (Hazardous Goods Information System) is developing a system to provide information about the movement and the handling of hazardous goods. There are a variety of regulations and standards in existence for different transport modes and different countries; these will be collated in an information base to provide a service which is multilingual for both the interrogation and presentation of information. Havre Port Innovation in Le Havre, France, is the project leader.

LINGUANET (TestBed LinguaNet), coordinated by ProLingua Ltd. Cambridge,

England, is developing a multilingual communications system for police departments in Europe. The initial system is based on an established user requirement and experience gained in communications between France and the UK concerning channel tunnel operations. It will use the controlled composition, manipulation, and exchange of text in different languages according to user specified templates.

TREE (Trans European Employment) is developing a service which will enable employers and employment agencies to publish details of job opportunities in a number of languages on the World Wide Web. Job seekers will then be able to view these details in their mother languages. Initially the service will be offered to the Leisure and Tourist industries where there is a great deal of labour mobility and where knowledge of the local language is often not a requirement. Again, Mari Computer Systems Ltd. is the project coordinator.

*4. Other relevant projects*

To complete the overview, which of course can only give a snapshot of what is ongoing in the R&D area of MT, we will mention some projects, which are not executed under the LE programme of the European Commission. I have selected the projects according to their relevance for stimulating future developments in the MT field.

The recently finished TRANSTERM project (LRE2 Actionline of the European Commission) has developed a portable toolbox for handling terminologies in a structured way with the ability to linking them to linguistic knowledge in the GENELEX format. The prototype of the project will be further developed by GSI-Erli within the Aleth product line (AlethGT -- gestion terminologique). Since TRANSTERM has developed a kind of terminology interchange format, discussions are underway with the MARTIF (Machine Readable Terminology Interchange Format) ISO-37 consortium for a harmonised joint interchange format.

Under the lead of IAI Saarbrücken the LSGRAM project (LRE1/2 Actionline of the European Commission) has developed language resources (analysis grammars and lexicons) for all EU member states based on the ALEP (Advanced Language Engineering Platform) framework. The project has demonstrated the maturity of ALEP for large-scale implementations and the flexibility of the ALEP system for the integration of different preprocessing utilities such as text handling. The work implemented in LSGRAM will be further investigated in the yet to be started ESPRIT project MELISSA, which aims to facilitate multilingual access to information systems. MELISSA will be coordinated by Software AG in Madrid, Spain.

MULTILINT (Multilingual Documentation and Document Production with Linguistic Intelligence) is a joint effort between BMW AG Munich and IAI Saarbrücken, Germany, for the integration of language technology and information technology in the area of multilingual technical documentation. The three-year project (1995-1998) is partly founded by the German Ministry of Commerce (BMWi --Bundeswirtschaftsminiterium). The application area is multifaceted and ranges from spell checking and terminology consistency checking to fully automated machine translation. One area of specialisation is a online multilingual information system for car repair and maintenance operations, where, in addition to the above technologies, Web technology plays a significant role in the integration process.

*5. Conclusions and Perspectives*

The current trend is for multilingual services based on networks (Internet, WWW, Intranets) to create new methods and methodologies for handling multilingual information, as well as new hardware and software solutions and the appropriate platforms.

This is of relevance for MT in three dimensions:

(1) standardised resources which are shareable among different MT suppliers,

(2) effective and efficient messaging protocols which may support several kinds of distributed online MT services (cf. my presentation at the EAMT workshop, Vienna August 1996),

(3) new methods and methodologies for MT applications.

The future of MT will be based on distributed agent-based MT system architectures which will supersede the traditional monolithic systems.

*6. Resources and References*

The Commission has made available detailed project descriptions of the LE1 projects at URL http://www2.echo.lu/langeng/en/le1/le1.html.

A further source of information is the proceedings of the EAMT Workshop in Vienna, Austria (August 29-30, 1996).

---

## A Machine Translation System for Minority Languages

*John Gareth Evans*

A machine translation system specifically suited for minority languages is being developed at the Swansea Institute of Higher Education. The overall system has a modular design and each module is in a different phase of development. It is hoped that a working prototype will be available for demonstration by the end of the year.

For the purpose of this communication, a minority language may be considered to a language which has not received and is unlikely to receive large amounts of funding in the near future for the development of MT systems. Examples would be Welsh, Romanian, etc. The prototype system is being developed using English-French simply because there is experience available at Swansea Institute in these two languages.

Emphasis is being placed on the development of generic components where language specific properties (details of grammar, vocabulary, semantics etc.) are held in separate data files from the executable program files. The processing of text in a source language results in the production of an intermediate, language-independent, unambiguous inter-lingua file. This will then be the basis of a target language generator.

The language dependent files (dictionaries, grammar definitions etc) can be produced manually, automatically generated from existing dictionaries in machine-readable form, or developed from parallel corpora. Although test files have been developed manually, the two latter approaches will be applied.

The philosophy is that for some applications of MT, e.g. the preparation of first drafts in specialist book translation, a lower quality translation is better than no translation at all. It is not intended that the system produce translations of the quality expected of tailored specific language-pair transfer systems, but rather that MT systems can be developed relatively quickly and economically where otherwise no system would exist.

For more information, contact: J Gareth Evans, Faculty of Computing, Swansea Institute of Higher Education, Mount Pleasant, Swansea SA1 6ED. Tel: (01792) 481144. E-mail: g.evans@sihe.ac.uk.

# TREE: TRans European Employment

TREE will provide a multi-lingual employment service via Internet. Employment adverts will be accepted electronically in several major community languages. Using advanced information analysis techniques, these will be stored into sophisticated, language independent, templates. A goal directed generator will then transform these templates as required into a variety of output languages. The generator will match the level of detail in the adverts to that of the user's query.

TREE will emphasise user requirements. The Consortium includes
MANPOWER Europe, and the Flemish Public Employment Service (VDAB). TREE is a project within the Language Engineering sector of the CEC's Fourth Framework Programme.

*Project Description*

One of the prerequisites for the realisation of an open European labour market is the accessibility of information about employment opportunities, both from the point of view of people seeking work and from the point of view of their potential employers. Today many EU citizens are denied full access to employment opportunities because (i) information about job opportunities may not be easily available, and (ii) even where it is available, it may not be available in his/her language. TREE seeks to address this problem by using language engineering and telematic techniques to enable multi-lingual access to information about job opportunities in the European context. The users are in the first place EU citizens in need of employment and prepared to move to a different part of the EU. Alternatively, users could be members of linguistic minorities already resident in EU states. Their foreign language skills are not sufficiently developed so as to make browsing job adverts in other languages easy. Typically users could range from unemployed Italian pizza cooks who would like to work in Sweden, British art students who would like to teach English in Rome or indeed members of ethnic groups who are disadvantaged by their cultural and linguistic diversity.

The users would be able to use the service telematically, via public terminals or information kiosks. The user interface should guarantee easy browsing and multi-lingual access to job adverts. Infrastructure support would be given by sponsoring local authorities and employment services. Initially, employment opportunities would come from the local authorities who are both sponsoring users, and major employers in their own right. As the service became known private companies will pay to advertise their employment opportunities on the service.

Through the combination of (i) public information service (through terminals or kiosks), (ii) multi-lingual database access, and (iii) easy-to-use browsing tools, this application will give more EU citizens access to the open European labour market.

*Technical Approach*

*The Internet*: The project would be based on the Internet (possibly indirectly), and users would be able to access it as a teleservice, or by using publicly accessible kiosks.

*Restricted Domain Machine Translation:* Employment adverts will be accepted electronically in several major community languages. Using advanced information analysis techniques, these will be stored into sophisticated, language independent, templates. A goal directed generator will then transform these templates as required into a variety of major and minor community languages. The generator will match the level of detail in the adverts to that of the user's query.

*Browsing*: The project will stress ease of use. An easy to use interface will be developed in conjunction with service end users and in accordance with their requirements. This will customisable (localised) for the various output languages.

---

## Oxford University Press active in Language Engineering

In July 1996 the Oxford University Press (OUP) announced the formation of the Oxford Centre of Language Resources to provide language expertise to the language engineering industry. The centre will have access to its own dictionaries and lexicographers, and to the British National Corpus containing 100 million words of contemporary British English. Its clients will be IT companies investing in and developing systems for speech recognition, content analysis, speech synthesis, and machine translation.

On the same day, OUP announced a joint project with Philips Dictation Systems to market a speech-to-text converter for PCs, enabling automatic production of written texts by direct dictation "naturally and fluently". OUP will be providing the information on language structure and use to support Philips' expertise in speech recognition technology.

---

# CORPORA and SERVICES

### Multilingual Evaluation Tool from TSNLP

*Lorna Balkan*

We would like to draw your attention to a multilingual evaluation tool that is now available. It consists of a database of test suites for English, French and German that have been constructed for evaluating Natural Language Processing Systems, but which may be useful for other purposes. The database consists of over 14,000 examples in English, French and German, which have been very systematically constructed with detailed annotations about various grammatical and other information. The test suites, support software, user documentation, and background documentation are available free from:

http://tsnlp.dfki.uni-sb.de/tsnlp/ (WWW)
tsnlp.dfki.uni-sb.de/tsnlp/ (anonymous ftp)

They have been produced by the University of Essex (UK), ISSCO (Switzerland), Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI) (Germany), and Aérospatiale (France) as part of LRE Project 62-089, Test Suites for Natural Language Processing (TSNLP).

*User Manual*

Major project results are documented in a user manual, which provides a description of the test data, the methodology which underlies their construction, and the tools which were developed in the project to aid test suite construction and use. The user manual is available in three volumes:

Volume 1: Background, methodology, customisation and testing. A description of the background to TSNLP, the methodology that underlies the TSNLP data, how the data can be customised, and how the data was used in practice to test a grammar checker.

Volume 2:Core Test suite technology. A description of the annotation scheme on which

the data is constructed, plus a description of the database (ANSI C and Access), and the test suite construction tool.

Volume 2b: Test suite technology. A description of the automatic test suite generation tool, and the lexical replacement tool

Volume 3: Test Data documentation. A detailed description of the data in the database.

TSNLP results are being made available free of charge to the academic and industrial community, in order that they might be widely used and accepted as an evaluation tool. Users are encouraged to report back any comments or criticisms. They are also encouraged to offer any extensions they make for inclusion in possible future releases of the database. The consortium is also happy to discuss producing customised test suites for specific users, at a cost to be arranged, and developing some of the construction tools.

*Contact point*: Lorna Balkan, CL/MT Group, Department of Language and Linguistics, University of Essex, Colchester C04 3SQ, UK (Tel: +44 1206 872092; Fax: +44 1206 872085; Email: balka@essex.ac.uk)

Project results will also shortly be available from the European Language Resources Association (ELRA), 87 Avenue d'Italie, 75013 Paris, France (Tel: +33 1 45 86 53 00; Fax. +33 1 45 86 44 88; Email: elra@calvanet.calvacom.fr; WWW: http://www.icp.grenet.fr/ELRA/home.html)

---

## European language corpora

*Ciara Walsh*

[From CORPORA List]

At http://www.ids-mannheim.de/telri/telri.html of the Trans-European Language Resources Infrastructre you can find information about resources at the ids and a number of partner institutions. Particularly good for eastern european resources.

At http://www.ruf.rice/edu/~barlow/corpus.html you will find a page maintained by Michael Barlow. This gives information on corpus resources in many languages, as well as software resources, bibliographies and pointers to other useful sites.

The ACL SIGLEX at http://www.clres.com/siglex.html a wide range of resources including electronic dictionaries, corpora and treebanks. However, much of this information is still under construction.

Finally, the UCREL homepage at the University of Lancaster http://www.comp.lancs.ac.uk/computing/research/ucrel has an excellent list of resources for anyone wanting to know what's available for corpus research.

---

## Cyc Upper Ontology now on World Wide Web

Cycorp, Austin, Texas, is the developer of the Cyc system for mechanized common sense knowledge. Cycorp has put its "Cyc Upper Ontology" on the Cycorp World Wide Web homepage at: http://www.cyc.com

This includes approximately 3000 of Cyc's highest concepts with the hierarchical links between them. It includes several hundred of Cyc's thousands of semantic relations, each with its argument-types (type signature). (The IF-THEN rules, part of Cyc's huge "expert system for the world in general," are not included in the current release. This release is a beta-test version of just the ontology hierarchy, with definitions and links.)

We invite others to build on and use the Cyc Upper Ontology (under liberal licensing terms, at no cost) in these areas and others we haven't imagined. Contact: Cycorp, 3500 W. Balcones Center Dr., Austin, TX 78759

---

## Dutch text corpus available

The Instituut voor Nederlandse Lexicologie offers you the possibility to consult a Dutch text corpus of ca. 38 million words, by the international computer network (Internet). In 1994 and 1995, a 5 Million Words Corpus with diversified composition and a 27 Million Words Newspaper Corpus have been made accessible in a similar way. Access is for free for non-commercial purposes.

The 38 Million Words Corpus 1996 consists of three main components: a component with varied composition (1970-1989), a newspaper component (Meppeler Courant, 1992-1995) and a legal component (1814-1989).

For additional information: e-mail to Helpdesk@Rulxho.Leidenuniv.NL.

---

## Linguistic Data Consortium

### New Releases

**Collins English Dictionary**

The Collins English Dictionary is available on CD-ROM from the Linguistic Data Consortium, and the disc contains the text from the typesetting tape as well as a parsed version of the text.

**Acoustic-Phonetic Continuous Speech Corpus: Far Field Microphone Recordings (FFMTIMIT)**

The FFMTIMIT corpus contains the previously-unreleased secondary microphone waveforms for the TIMIT Acoustic-Phonetic Continuous Speech corpus. The primary microphone waveforms are available from the LDC on NIST Speech Disc 1-1.1 (LDC93S1).

**Continuous Speech Recognition Corpus-IV (Hub-3)**

This set of CD-ROMs contains all of the speech data provided to sites participating in the DARPA CSR November 1995 Hub-3 Mulit-Microphone tests.

**RM1: The Resource Management-Word Data Continuous Speech Database: Isolated and Spelled Word Data**

This CD-ROM contains previously-unreleased isolated-word and spell-mode (spelled out words) speech data from the (D)ARPA Resource Management (RM1) Corpus.

**DCIEM Sleep Deprivation Study: Map Task Dialogues**

This set of CD-ROMs contains the materials used to collect all 216 spoken dialogues digital audio, orthographic transcriptions, documentation, and source code for tools. The dialogues were selected to provide balanced representation at different points in a sleep deprivation experiment.

**LDC-Online**

LDC-Online is a new search and retrieval service, offering convenient WWW access to the text and speech corpora of the Linguistic Data Consortium (LDC). For more detailed information, or to try it out, see the LDC-Online item on the LDC's home page (http://www.ldc.upenn.edu). Information about LDC is also available via ftp at ftp.cis.upenn.edu under pub/ldc; for ftp access, please use "anonymous" as your login name, and give your email address when asked for password.

---

## Moby available from ILASH, Sheffield

[From LINGUIST List, Vol-7-1045]
ILASH have announced that the fruits of the Moby project are being placed in the public domain. The files comprising the project are now available from

ftp://ftp.dcs.shef.ac.uk/share/ilash/Moby/

either as a complete distribution [26MB] or as set of subprojects: Moby Hyphenator (185,000 entries fully hyphenated); Moby Language (Word lists in five of the world's great languages); Moby Part-of-Speech (230,000 entries fully described by part(s) of speech, listed in priority order); Moby Pronunciator (175,000 entries fully International Phonetic Alphabet coded); Moby Shakespeare (The complete unabridged works of Shakespeare); Moby Thesaurus (30,000 root words, 2.5 million synonyms and related words); Moby Words (610,000+ words and phrases. The largest word list in the world). For more details, see:

http://www.dcs.shef.ac.uk/research/ilash/Moby/

---

## INTERSECT corpus available on Web

INTERSECT, a parallel corpus project at the University of Brighton, England, now has a Web site. The URL is:

http://bmserver.bus.bton.ac.uk/FormatGraphical-Normal/BusSchool/Research/LangCent/Intersect.html

Further information: Raphael Salkie, The Language Centre, University of Brighton, Falmer, Brighton, BN1 9PH, England. (Tel: +44 1273 643335; Fax: +44 1273 690710; Email: r.m.salkie@brighton.ac.uk)

---

## Information on Controlled Languages

*Christina K. Alexandris*

[From LINGUIST List, Vol-7-923]

For anybody interested on obtaining information on controlled languages, the following websites and mailing lists can be helpful
WEB SITES :
http://wwwots.let.ruu.nl/Controlled-languages
http://wwwots.let.ruu.nl/Controlled-languages/data/references.html
http://hfskyway.com/mtng10/drury.html
-in the latter website , a recent study of AECMA Simplified English
called "Field Evaluation of Simplified English Workcards" is available
http://www.ccl.kleuven.ac.be/cgi-bin/seccdemo.cgi

-one can run the SECC controlled language at this web site
MAILING LISTS :
Send a message to: majordomo@let.ruu.nl
containing the following text only:
subscribe controlled-languages <your email address>
WORKSHOPS :
The Proceedings of the workshop on controlled languages (Univ. of Leuven,Belgium [see MTNI#14] can be obtained if you send email to Geert Adriaens (geert.adriaens@csl.sni.be).

---

## MtRecode - Character conversion program from MULTEXT

*Jean Véronis*

MtRecode is a program for translation between various character sets, developed in the framework of the MULTEXT project. It has some of the functionality of the GNU `recode' tool, but it is based on different principles and is oriented towards SGML text manipulation. ISO 10646 is used internally as a pivot in the character translation process. When exact translation into a character is not possible, MtRecode can use SGML entities as a fallback. Conversely, MtRecode understands SGML entities in the input and can recode them into characters of the target character sets, if they exist. MtRecode is completely customizable: the user can add new character sets and/or entities by providing tables that map characters and entities to ISO 10646.

C source code for UN*X platforms and documentation can be freely downloaded for non-commercial, non-military use (see our user agreement) from:

http://www.lpl.univ-aix.fr/projects/multext/MtRecode/

Note that MtRecode is an alpha version with (probable) bugs and limitations. It is being distributed "as is" in order to solicit feedback. We invite the user community to send comments and advice, provide additional character tables, etc.

---

## Software from Lingsoft, Inc.

This Helsinki, Finland software company which develops linguistics software for text retrieval and information management systems, specializes in the processing of English, German, Swedish, Finnish. Also Estonian, Russian, Swahili. Coming: French, Italian, Norwegian.

Examples: ENGTWOL (English morphological analysis) -- Contains 75,000 base forms, recognizing over 300,000 word forms; ENGCG (English Constraint Grammar) -- A general surface syntactic constraint grammar with a full functional tag set for English (correctness 99.7%, ambiguity left 3-7%).

The products are available for academic and non-profit research. Contact: Markku Norberg (Markku.Norberg@lingsoft.fi), Lingsoft Inc., Museokatu 18 A, 00100 Helsinki, Finland (Tel. +350 9 499 556; Fax +358 9 440 602; WWW: http://www.lingsoft.fi/)

---

## CLEARS: an educational and research tool for computational semantics

[From LINGUIST List, Vol-7-1150]

The CLEARS tool provides a graphical interface allowing interactive construction of semantic representations in a variety of different formalisms, and using several construction methods. CLEARS (Computational Linguistics Education and Research for Semantics) was

developed as part of the FraCaS project which was designed to encourage convergence between different semantic formalisms, such as Montague-Grammar, DRT, and Situation Semantics.

CLEARS is freely available on the WWW from the following address, which contains further documentation: http://www.coli.uni-sb.de/~clears/clears.html. Further information: clears@coli.uni-sb.de

---

### Language Conference List

[From Linguist list 7-992]

The list located on the WWW at URL http://www.clark.net/pub/royfc/confer.html includes conferences for anyone interested in any aspect of natural language: linguists, translators, interpreters, teachers of languages, those who are involved in natural language processing, et al. Changes, updates, corrections or comments via e-mail to royfcoch@clark.net.

---

# PUBLICATIONS ANNOUNCED

### Machine Translation

*Changes in MT journal*

The journal "Machine Translation" has a new editor and editorial board.

*Editor*: Harold Somers, Centre for Computational Linguistics, UMIST, Manchester

*Book Review Editor*: Doug Arnold, Department of Language & Linguistics, University of Essex, Colchester

While respecting its historical link with the field of MT, "Machine Translation" is changing and broadening its scope of interest to encompass all branches of Computational Linguistics and Language Engineering, wherever they incorporate a *multilingual* aspect. We therefore welcome submissions to the journal on *theoretical, descriptive or computational aspects* of any of the following topics:

* machine translation and machine-aided translation
* human translation theory and practise
* multilingual text composition and generation
* multilingual information retrieval and natural language interfaces
* multilingual dialogue systems and message understanding systems
* corpus-based and statistical language modelling
* connectionist approaches to translation
* compilation and use of bi- and multilingual corpora
* discourse phenomena and (human or machine) translation
* contrastive linguistics
* software localization and internationalization
* speech processing, especially for speech translation
* computational implications of non-Roman character sets
* language engineering for minority languages
* history of machine translation

We would also welcome your suggestions about other features you would like to see in this journal, for example, special issues, squibs, topical comment. See our web pages at

http://www.ccl.umist.ac.uk/harold/MTjnl

*Submission of manuscripts*: For information about submission, contact the Editorial office, Machine Translation, Kluwer Academic Publishers, PO Box 17, NL-3300 AA Dordrecht, The Netherlands; tel. (+31) 78 639 2911; fax 2254; e-mail editdept@wkap.nl; From North America contact Machine Translation Editorial office, PO Box 230, Accord, MA 02018-0230.

Any other comments or queries to Harold Somers (MT journal), Centre for Computational Linguistics, UMIST, PO Box 88, Manchester M60 1QD; fax +44 161 200 3099; email harold@ccl.umist.ac.uk

---

## Natural Language Engineering

**Special Issue on**
**Knowledge Representation for Natural Language Processing**
**in Implemented Systems**

This special issue is intended to be a forum for the presentation of the state-of-the-art in implemented knowledge representation and reasoning (KRR) systems for general natural language processing (NLP). We are interested in papers that address or describe implemented knowledge representation systems that facilitate natural language processing for implemented systems.

Deadline for submissions is December 31, 1996. Further information: http://tigger.cs.uwm.edu/~syali/jnle-kr-nlp/

---

## Twente Workshops on Language Technology

**Proceedings Available**

Volumes in the TWLT Proceedings Series can be ordered from the secretariat, at the address below, by mail, fax or email. For previous volumes, ask the secretariat. The price of each TWLT Proceedings volume is NLG 50,- + VAT (17,50%) + administration and mailing costs.

Recent Proceedings: TWLT 11, June, 1996 (Dialogue Management in Natural Language Systems); TWLT 12, September, 1996 (Computational Humor: Automatic Interpretation and Generation of Verbal Humor.)

SETI secretariat, University of Twente, Department of Computer Science, P.O. Box 217, 7500 AE Enschede, The Netherlands (Tel: +31 53 4893680; Fax: +31 53 4893503; Email: twlt_secr@cs.utwente.nl)

---

## Survey of the State of the Art of Human Language Technology

This book now available (http://www.cse.ogi.edu/CSLU/HLTsurvey/) is a survey consisting of articles by 97 authors in the following chapters: Spoken Language Input, Written Language Input, Language Analysis and Understanding, Language Generation, Spoken Output Technologies, Discourse and Dialogue, Document Processing, Multilinguality, Multimodality, Transmission and Storage, Mathematical Methods, Language Resources, Evaluation.

Within a few months, the Survey will be published as a book by Giardini Publishers in Italy and by Cambridge University Press elsewhere. The electronic version of the Survey will remain on-line, but will be modified slightly based on copy-editing by the publishers.

The Survey was funded by the National Science Foundation and the European Commission, with additional support provided by the Center for Spoken Language Understanding at the Oregon Graduate Institute and the University of Pisa.

---

## International Journal of Corpus Linguistics

IJCL presents a wide range of views on the role of corpus linguistics in language research, lexicography and natural language processing (NLP). IJCL seeks to publish research that views language as a social phenomenon that can be investigated empirically on the basis of authentic spoken and written texts. Corpus linguistics specifies corpus design in respect to research interests, provides computational methods of extracting linguistic knowledge, and conceives tools to validate the accuracy of linguistic description.

IJCL aims to conciliate the expectations of language industry with the goals of academic research. Corpora are the basic resources in language engineering. It is the linguistic knowledge extracted from corpora that determines the performance on any NLP application. IJCL is a forum to exchange and share expertise, visions as well as information on resources and tools.

IJCL invites relevant contributions. For information on contributions and guidelines contact: Dr. Wolfgang Teubert, Institut für deutsche Sprache, Postfach 10 16 21, D-68016 Mannheim, Germany (Fax: +49 621 1581 415; Email: IJCL@ids-mannheim.de)

For information on subscription/subscription orders contact: John Benjamins, PO Box 75577, 1070 AN Amsterdam, The Netherlands; or: John Benjamins North America Inc., PO Box 27519, Philadelphia PA 19118 0519, USA.

---

## The Finite String

*Richard Sproat*

[From LINGUIST List, Vol-7-494]

The FINITE STRING is the quarterly newsletter of the Association for Computational Linguistics (ACL), and is a supplement to the journal *Computational Linguistics*. The function of the STRING is to provide a forum for advertising events -- especially conferences and workshops -- of interest to the ACL community.

Starting in 1996, the STRING is published as an html document, available at the ACL website at Columbia University. The URL is:

http://www.cs.columbia.edu/~acl/finstring.html

There is a European mirror at:

http://issco-www.unige.ch/eacl/acl/finstring.html

Organizers of conferences and workshops that are likely to be of interest to the ACL membership are invited to submit a description of the conference to the STRING. Submissions must be sent via email to the Finite String editor. No hardcopy submissions will be accepted. Submissions should include the "Subject:" line "Finite String Submission".

Richard Sproat, Editor, FINITE STRING, Speech Synthesis Research Department, Bell Laboratories, Lucent Technologies (Email: rws@research.att.com)

---

## LINGUIST List on WWW in Europe

*Graham Katz*

[From LINGUIST List, Vol-7-781]

The complete contents of the LINGUIST network, including the most recent LINGUIST postings, is now available at the following address: http://www.sfs.nphil.uni-tuebingen.de/linguist/

Graham Katz (Seminar für Sprachwissenshaft, Universität Tübingen; email: katz@sfs.nphil.uni-tuebingen.de)

---

## Survey of MT Teaching

*Resources and Methods for teaching Machine Translation: A Survey*

We are interested in hearing from anyone who is involved in the teaching of Machine Translation (MT). We are conducting a survey of resources and methods used in the teaching of MT and are particularly interested to hear from those who use, or would be interested in using, practical MT systems as a teaching aid. The survey is partly sponsored by ELSNET.

We are interested in your ideas, experience, needs, etc. etc., but in the first place we are interested in knowing WHO YOU ARE. Particular questions we will be addressing include what demand there is for practical MT systems as a teaching aid, what commercial MT systems are presently available, at what cost, etc. The report and other results of the survey will be sent to everyone who replies. Concrete outcomes may include some ideas about what how courses can be structured, what teaching aids are available. If resources permit, we may try to organize a workshop for face to face exchange of ideas.

Reply to: Doug Arnold (doug@essex.ac.uk; Tel: +44 1206 872084); Lorna Balkan (balka@essex.ac.uk; Tel: +44 1206 872092); Louisa Sadler (louisa@essex.ac.uk; Tel: +44 1206 872082), CL/MT Group, Dept. of Language & Linguistics, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, UK. (Fax: +44 1206 872085; WWW: http://clwww.essex.ac.uk/)

---

## Wanted: Examples of English-French MT

*Louise Brunette*

[From LINGUIST List, Vol-7-933]

I am looking for moderately technical texts translated by MT.
Even if there is a lot said on MT, it is difficult to find MT translated texts in Canada and especially in Quebec. The purpose of my request is to get some material from which I could teach postediting to my undergraduate students in the framework of a course entitled : Revision of French translations. Weather broadcasts are of no help since they are hardly revised.

Contact: Louise Brunette, Concordia University, Montreal (brunettl@ERE.UMontreal.CA)

---

# PUBLICATIONS RECEIVED

*Journals*

**AAMT Journal** *no.14, March 1996; no.15, June 1996; no.16, September 1996.* In Japanese only.

**ALPNET news** *issue 2, 1996.*

**Elsnews** *vol.5 no.2 (May 1996)*. Contents include: Speech analysis systems (Dawn Griesbach). -- Computational linguistics in Malta (Michael Rosner).

**Language International** *vol.8 no.3 (June 1996).* Contents include: TERMBASE for Microsoft Windows (Magdalene Clegg). -- Translation Assistant. -- The 1996 EAMT Workshop (Colin Brace); *vol.8 no.4 (August 1996)*. Contents include: Lingo 1.2: language assistant for Windows 3.1 (Magdalena Clegg). -- The living dictionary (Bob Clark) [Logos Dictionary on-line]. -- WWW Japanese translation [LogoVista E to J Internet].

**LISA Forum Newsletter** *vol.5 no.2 (May 1996)*. Contents include: Translation and the Internet (Robin Bonthrone). -- The Spanish language: dream or nightmare? (Javier Garcia Alvarez). -- Teletranslation - a brave new world (Zeger Karssen). *vol.5 no.3 (September 1996)*. Contents include: Screams from the quality jungle (Robin Bonthrone). -- Present and future needs in the CAT world (Matthias Heyn). -- EUTERPE on an intranet (Cornelis van der Laan and Matthias Heyn). -- Getting our act together? (Deborah Fry) [on proposed European Association for Terminology.]

**Literary and Linguistic Computing** *vol.11 no.2 (June 1996)* Contents include: An algorithm for generating a dictionary of Japanese scientific terms (Y.-S.Kang and A.A.Maciejewski). -- A statistical learning approach to improve the accuracy of Chinese word segmentation (C.-H.Leung and W.-K.Kan). *vol.11 no.3 (September 1996)* Contents include: An upper bound estimate for the entropy of Korean texts (Y.S.Han et al.)

**Machine Translation Review**: the periodical of the Natural Language Translation Specialist Group of the British Computer Society, *issue no.3 (April 1996)*. Contents include: Implementing an efficient compact parser for a machine translation system (J.Gareth Evans). -- Using Icon for text processing (David Quinn). -- Book reviews.

**Natural Language Engineering** *vol.1 part 4 (December 1995)*. Contents: High speed feature unification and parsing (John C.Brown). -- Engineering 'word experts' for word disambiguation (Daniel Berleant). -- POETIC: a system for gathering and disseminating traffic information (R.Evans et al.) -- Book reviews.

*Books*

Kugler, M.; Ahmad, K.; Thurmair, G. (eds.) **Translator's workbench: tools and terminology for translation and text processing.** (Research Reports ESPRIT, Project 2315.) Berlin: Springer, 1995. ix,181 pp. ISBN: 3-540-57645-2

O'Hagan, Minako. **The coming industry of teletranslation: overcoming communication barriers through telecommunication**. Clevedon/ Philadelphia/ Adelaide: Multilingual Matters,

1996. xiii, 120 pp. ISBN: 1-85359-326-5, 1-85359-325-7 (Topics in Translation 4)

*Conference proceedings*

Steffens, Petra (ed.) **Machine translation and the lexicon**. Third International EAMT Workshop, Heidelberg, Germany, April 1993. Proceedings. Heidelberg: Springer, 1995. x,251pp. ISBN: 3-540-59040-4 (Lecture Notes in Artificial Intelligence 898)

**Le traitement automatique du langage et les applications industrielles, Natural language processing and industrial applications**. NLP+IA/TAL+AI, 4-6 juin 1996, Moncton, N.-B. Canada. Moncton: Université de Moncton, GRÉTAL, 1996. 2 vols.

Galinski, Christian; Schmitz, Klaus-Dirk (eds.) **TKE'96: Terminology and knowledge engineering.** Proceedings of the Fourth International Congress on Terminology and Knowledge Engineering, 26-28 August 1996, Vienna, Austria. Frankfurt/M: Indeks Verlag, 1996. viii, 461 pp. ISBN: 3-88672-207-4

EAMT Machine Translation Workshop: TKE'96, Vienna, Austria, 29-30 August 1996. Programme and abstracts. [Luxembourg: 1996]

---

*Items for inclusion in the 'Publications Received' section should be sent to the Editor-in-Chief at the address given on the front page. Attention is drawn to the resolution of the IAMT General Assembly, which asks all members to send copies of all their publications within one year of publication.*