MT News International

Newsletter of the International Association for Machine Translation

ISSN 0965-5476 Issue no.8, May 1994

IN THIS ISSUE:

Editor-in-Chief:
John Hutchins, The Library, University of East Anglia, Norwich NR4 7TJ, United Kingdom
Fax: +44 (603) 259490; Email: J.Hutchins@uea.ac.uk; 100113.1257@compuserve.com
Regional editors:
AMTA: Joseph E.Pentheroudakis, Microsoft Corporation, Redmond, WA 98052, USA
Fax: +1 (206) 936-7329; Email: josephp@microsoft.com
EAMT: Tom C.Gerhardt, CRPCU/CRETA, 13 rue de Bragance, L-1255 Luxembourg.
Fax: +352 (44) 73 52; Email: tom@crpculu.lu
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 Americas, with the generous assistance of Microsoft Corporation

NEWS of SYSTEMS and PRODUCTS

Danish patent system based on Eurotra experience

Colin Brace

[From *Language Industry Monitor*, no.19, pp.1-4.]

Eurotra may have never replaced Systran, but its ghost lives on in Denmark, where a brand new MT system based partly on Eurotra work recently went into operation.

This past December saw the installation of a new English-to-Danish machine translation system at Lingtech, a translation company based in Copenhagen, specializing in translating patents. The new system, called **PaTrans**, was developed by the Center for Sprogteknologi (CST) -

previously Eurotra Denmark. The development of PaTrans marks a new era for post-Eurotra MT in Europe and is a splendid achievement for the CST. Under the leadership of director Bente Maegaard, the CST, predominantly research-oriented in the past, has delivered a functional MT system to specification, on time, and within budget.

To understand the precise nature of this success, it should be noted exactly what PaTrans is. This English-to-Danish system has been designed for the domain of patents, initially petrochemical patents. It is decidedly not a general-purpose system, and consequently lacks the kind of facilities and resources that a shrink-wrapped package would be expected to have; its users are also prepared to both pre-edit the input and post-edit the output of this Unix-based system. But the CST's customer is getting what it wants. Lingtech director Viggo Hansen says that PaTrans satisfies all the design and performance criteria that were established at the beginning of the project. Previous to the installation at Lingtech, the software ran for several months at the CST to iron out any remaining wrinkles. When the system was delivered in December, there were simply no surprises. "You're a brave man," we remarked recently to Mr Hansen, who replied with a smile, "many people say that." Hansen's confidence has clearly been rewarded; Lingtech's MT gamble has paid off.

The inevitable question arises: How much of PaTrans is based on Eurotra? "It depends on who is asking," quips project manager Annelise Bech. Commission technology boosters would obviously love to hear that the system is a direct spinoff of the Eurotra program, while Eurotra insiders would shudder at the thought of *production* software based too closely on that unwieldy collective effort. In the final tally, the question is virtually impossible to answer; moreover, it is not terribly relevant. While you might be able to count lines of code or lexicon entries, how do you quantify the collective experiences of a group of researchers over ten year's time? Maybe Eurotra demonstrated how not to do certain things. The CST feat, then, is not that it somehow recycled Eurotra technology, thereby putting something back in the communal kitty, but that it has been able to capitalize on the human resources cultivated during the Eurotra years. And that, in hindsight, makes the Eurotra exercise seem at least partly worthwhile.

Inevitably, the CST has drawn on past work, benefitting, points out Bech, from the formal grammatical descriptions the group previously developed for Eurotra. But whereas Eurotra was by definition designed to be as general as possible to accommodate all potential language pairs, PaTrans was expressly designed and heavily optimized for just one language pair, going in only one direction. Among other things, that means parsing times can now be measured in seconds, not minutes, as was the case with Eurotra.

Speed is one consideration in a production system; robustness is another. PaTrans includes what Bech calls a *failsafe* strategy. This encompasses not only morphology-based heuristics for unknown words, but also the robust parser that does not plead "no parse," even when dealing with ungrammatical input. "The system does not require the user to pre-edit input texts to conform to a controlled language," says Bech. "There are no restrictions on the linguistic constructs it will accept. However, its coverage definition does specify what the parser can deal with before it resorts to the failsoft strategy."

Additional robustness is provided by the PaTrans pre-parser, largely the work of the CST's Brad Music. Explains Music, "we can do some useful things in the pre-parsing stage, such as delineate sentence boundaries and lists and tag some constituent parts. We also identify some coordinate structures and prepositional phrases. The more you can trap at this stage, the better." The fewer the ambiguities, the less likelihood of the parser overspecifying, and this means fewer parse trees generated by the system's chart parser. And that, in turn, means better performance.

Reflecting the nature of the input material, PaTrans is largely syntax-based, performing, in the words of Bech, "deep syntactic analysis, like the Eurotra model." However, she adds that the PaTrans lexicons do include a few attributes to facilitate some translation ambiguities, e.g. there is a feature called *measure*. PaTrans also draws on semantic distinctions in the partitioning of lexical items in domain-specific term dictionaries, such as petrochemical, chemical, and mechanical engineering. "These partitions are physically distinct," explains Bech. "The items are actually labelled for domain." Accordingly, a PaTrans term-dictionary will be assigned a basic semantic category, e.g. composition (chemical), composition (legal), etc. While PaTrans's grammatical coverage may be limited to patent texts, control of its lexical coverage is squarely in the hands of the user.

Naturally, a lot more goes into a production MT system than just linguistics; you also need a document-handling component and end-user facilities, both of which are non-trivial development tasks. While the interface to the translation module was rather spartan when the CST was demonstrating the system in November, it will eventually be integrated into an "administrative" front-end. This will provide a menu-based interface for both the translation module and the pre- and post-editing facilities. It will also offer some document-handling tools, such as an archiving system for the patent texts. However, in terms of the user-interface, the highest priority has been the term- and dictionary-entry tool, called PaTerm. Jacketed in an elegant X/Windows interface, PaTerm offers two levels of support to end-users.

Level one is for the user who requires lots of guidance. He or she is taken through the process step by step, prompted a question at a time by on-screen examples. Level two is for the experienced user; there is only one screen and no examples. The user fills in the information in short form by clicking on boxes or using the keyboard. Level two also offers a *template* function for adding a batch of similar terms which differ only in lexical values. The coding tool is "intelligent," says Bech; it computes itself a number of values based on the input of the user. It is also very fast. The CST is quite proud of PaTerm; it designed and implemented the tool from scratch.

"It was absolutely essential that the PaTerm be easy to use," explains Viggo Hansen. "We wanted people with only a general knowledge of grammar to be able to use it." Lingtech had a person coding the PaTrans dictionary fulltime before the system was introduced at Lingtech; she spent several months coding terms in the chemical and petrochemical domains. PaTerm reflects the fact that not only linguistic expertise but also software engineering skills were required for the system; Bech says that programming the X/Windows interface posed more difficulties at given moments than developing and implementing the linguistic engine.

Lingtech was established several years ago by two large Danish patent attorney firms, Lehmann & Ree and Hofman-Bang & Boutard, which found themselves spending so much time on translation that they decided to join forces and establish a separate translation company. As in most countries, patents are valid in Denmark only when a Danish-language translation is registered locally. While patent translation may be one aspect of the job, a patents attorney's primary role is to assist inventors and researchers obtain patent protection and other kinds of industrial property rights, such as trademark and design protection. Viggo Hansen was originally hired as a consultant by the two firms to ascertain the feasibility of automating the patent translation process. This resulted in the newly reorganized CST performing a feasibility study in 1992 that led to the signing of a contract to develop a translation system.

Patents are legal documents and, as such, call for meticulous, literal translations. Is MT suitable for such an application? Hansen acknowledges that PaTrans translation requires both pre-editing and careful reviewing by an experienced translator, but the effort is worth it, although the actual benefits in terms of increased efficiency and accuracy can only be measured after the system has been in use for some time. "In a patent, each word has one and only one meaning. And the terminology is fixed." Hansen adds that these are fairly long documents, four thousand words on average - although patents of American origin tend to be even longer. "It would not be worthwhile running a one to two page text through such a system," he points out. The company already translates between one and two thousand patents a year, and the volume is growing, exceeding even projections supplied by the European Patent Office in Munich.

A stumbling block remains the fact that most of the English patent texts are supplied in hardcopy form; only one of Lingtech's customers regularly supplies electronic files. This implies the extra step of scanning in a text, not always a faultless process. Hansen acknowledges the problem, saying, "we *know* that each of these texts must be stored on a diskette *somewhere* in the world. It is just a matter of tracking it down." In the meantime, Lingtech provides an economic incentive - a discount - for patents supplied in electronic form.

Lingtech selected the petrochemical domain because it seemed a suitable subject for an initial foray into MT - and because its two parent companies do a lot of business in this field. But if everything goes well, petrochemicals could be just the start; hansen suggests that the system could also be extended to cover pharmaceuticals, electronic equipment and other domains. The goal, says Hansen, is "full coverage" of all the technical areas in which the parent offices are active. Further down the line, other kinds of technical documentation for third parties could also be a possibility. Whatever direction the Lingtech takes, it faces coding lots and lots of terminology. Says Hansen, "PaTerm is the key to the future expansion of the system. That's why we put so much effort into it."

Looking back on the two and half year trajectory of the project, Hansen has nothing but praise for the CST. "It is a highly competent group," he says. "There were no mishaps." But Hansen, a computer industry veteran, no doubt contributed substantially to the success of the endeavour himself by bringing to bear his previous experiences in large office automation projects. He notes that a thorough system analysis was performed before a line of code was written.

With the introduction of new MT systems a rare occurrence, PaTRans seems like a very promising paradigm for the present age. General-purpose MT is still an oxymoron. System like Systran, Logos, and Metal are largely used in narrow domains - and only after extensive lexicon work. Maybe it would be better to market MT systems in the form of applications like PaTrans, for at least then the substantial cost of "customization" would be up front.

In any event, PaTrans is more evidence - if you still need it -that the most interesting work in the field of MT *applications* is taking place in Europe, CMU notwithstanding. But why Denmark in particular? MT competence and patents can be found in many European countries. PaTRans seems to be the result of a happy confluence of two factors: an experienced researcher who had the formidable skills needed to pilot her team through the transition from research to development, and a customer who could formulate *exactly* what he wanted.

News from Winger92

[Adapted, with permission, from on article in *Language Industry Monitor*, no.19]

Danish software house Winger has released a set of electronic bilingual dictionaries for translators. The DOS-based package, called Book'92, is available currently for Danish-English, French-English and English-Spanish. The dictionaries, based partly on Collins dictionaries, share the same basic format as the Winger'92 dictionaries and can be run in conjunction with wordprocessing software with the usual hot-key and copy-paste functions. At present, the memory required is 60KB, but a future release is expected to reduce this to 10KB. Each language pair costs 400 Danish kronor (ca. US$100).

Winger has also developed SuperBook (DKr.1000), a dictionary administration package allowing the creation and maintenance of up to 64 subject-specific dictionaries across a network. It is seen as an excellent means of collecting data for Winger'92 dictionaries. New Book'92 entries can be merged easily into Winger'92 dictionaries, and the developers regard Book'92 as a way of gently initiating people into the technology of computer-aided translation. When they discover the productivity gains with simple dictionary lookup tools, they may be tempted to investigate the more ambitious Winger'92 package.

Winger'92 has now been on the market for 5 years. In addition to the original Danish-English language pairs, Winger also offers Spanish-English in both directions. The latest release includes improvements to the user interface during post-editing: deleting alternative target offerings, inversion of words in the target text, copying and pasting words to buffers. There is also multitasking capabilities, enabling users to work in the foreground in interactive mode, while the program translates other text in the background. Other recent enhancements include facilities for defining style-specific grammars, and in an extended version the possibility of direct access to the MT engine itself. In its current form, there is a rudimentary translation memory function suitable for updating different versions of documents.

Recently, the company has become involved in collaborative ventures. With the Center for Sprogteknologi (Copenhagen) - the creator of the PaTrans system described elsewhere in this issue - and with the IAI in Saarbrücken, Winger has embarked on the development of a Danish-German version, funded by the European Community. The CST will be responsible for transferring the Eurotra grammar into the Winger grammar programming language SALT, and Saarbrücken is preparing and testing the dictionary procedures. In another collaboration with the Danish wordprocessor company Dansk System Industri, Winger is working on a Danish-Russian office system encompassing a wordprocessor, a communications package, a database and a translation module.

The STYLUS LINGVO systems from Russia

[Based on publicity leaflet, translated by Evgeny Lovtsky]

The STYLUS and LINGVO systems are the joint development of two companies: BIT and PROekt MT. STYLUS is a system for translating between the language pairs Russian and English, capable of translating 8MB of information (4,000 pages) overnight in batch mode. It can be used for commercial proposals, business documents and technical documentation. LINGVO offers access to a range of specialist computer-based dictionaries, enabling translators to find words in several dictionaries simultaneously, and to insert chosen equivalents in translations. The following range of tools are available:

FINEREADER: an OCR reading both Cyrillic and Latin fonts (one typed page per minute on a 12 MHz PC AT)

LINGVO CORRECTOR: an automatic spelling corrector for Russian and English (100-150 words per second)

STYLUS: sentence-by-sentence or batch translation from English, German, French and Italian into Russian, and from Russian into English (5 pages per minute), in the fields of computer programming and business correspondence. (It is claimed that 90% of translation requires no post-editing.)

LINGVO: memory-resident computer dictionaries, with screen input, dictionary lookup, insertion into texts, with facilities for users to create dictionaries (requires 3.9KB RAM.) It includes the LINGVO program shell 4.0 (for DOS and Windows). The following dictionaries are available (with nos. of entries):

Bilingual (English/Russian): common vocabulary (38,000 entries), commerce (1,200), trade (4,000)

English-to-Russian: PC user (10,000), information science (5,600), computer hardware (35,000), economics (20,000), popular science (4,000), aviation (33,000), machine building (8,000), robotics (15,000), petroleum (7,700), mathematics (8,000)

Russian-to-English: economics (20,000), mathematics (12,000), polytechnic (15,000).

TransLand: Japanese-English Translation Software
A new product from Brother Industries Ltd.

[From *AAMT Journal no.5*; translated by Shravan Vasishth]

1. Background. In order to prepare documents to be sent to our overseas factories in the US and the UK, we have developed a document creation support system for mainframes and workstations.

This system helps to rapidly prepare documents required by overseas factories which are engaged in the production of appliances (such as office machines, electrical appliances, etc.) designed in Japan. We developed the Japanese-English translation support software in order to translate documents written in Japanese into English and this software is now being used as a part of the total support system.

The present product has an improved, more general-purpose grammar and dictionary. Moreover, the processing program has a higher speed and a smaller memory requirement, thus making it possible to use a PC.

It is common knowledge that the performance of the PC has been steadily improving, it has become cheaper and the number of non-specialist PC users is increasing. Today, PCs exist that can do processing which was earlier restricted to workstations.

The criteria we have followed in developing our product are as follows:

* A personal user should be able to use it easily

* The translation quality should be the same as or higher than that of workstation-type application software

* It should have a simple and easy user interface.

2. Special characteristics of the product.

* High translation quality

As mentioned above, software developed for mainframes and workstations can now be used on PCs without any deterioration in basic performance.

The translation method adopted is the 'meaning transfer' method.

Keeping processing speed and accuracy of description in mind, the grammar description language previously in use has been improved upon. Moreover, dictionary information has also been greatly enhanced. All this made it possible to transfer the software to PCs, retaining at the same time the high-quality translated output hitherto obtainable only through workstations.

* User dictionary registration

The parts of speech that can be entered into the user dictionary are: the verb *suru* ('to do'), nouns, verbs, adjectives (conjugative and non-conjugative), adverbs, conjunctions and classifiers.

When entering data into the user dictionary, the user must also enter any unusual grammatical and semantic information.

If the extent of information that can be entered by the user into the user dictionary were to be made too large, the system dictionary would be adversely affected, resulting in a possible deterioration of translation quality. To prevent this from happening, the amount of information that can be entered by the user has been set so as to neither inconvenience the user nor adversely affect the system dictionary.

Furthermore, user dictionary registration has been made easier by providing an online help while data is being entered.

* Simple editing and display functions

Display and editing functions for Japanese text and the translated English text are simple and have been kept to the minimum required so that even a beginner can start using these immediately. Post-editing, proof-reading, etc. are carried out by the user on word-processing DTP software that he is familiar with.

A basic operation is to read an MS-DOS text file, to translate it and produce output.

* Other features

TransLand runs on the main Japanese FEPs: ATOK 6, 7 and 8, WX II, Matsutake, and MS-KANJI are set automatically.

3. Future schedule.

We intend to develop TransLand further. Some of the improvements planned are:

- to make TransLand usable on other PCs and OSs.
- to develop dictionaries in specialized fields.

Specifications:

Operational environment:

Applicable models: NEC PC98 series

CPU: 80386SX or greater (80486 recommended)

Memory: protect memory 5MB minimum (8MB or more recommended), plus main memory 640KB
HDD: Free disk space 16MB minimum
OS: MS-DOS ver. 3.3x, 5.0x
Software specifications:
Dictionary: System dictionary approx. 43,000 words
User dictionary: Limited by memory
Translation method: Transfer-by-meaning method
Speed: (80486/33MHz) 5-30 seconds per sentence
(depending on complexity)
Operation mode:
one-sentence input and translation
MS-DOS text file input & one-sentence translation
MS-DOS text file translation
Output: MS-DOS Japanese text file
MS-DOS English text file
MS-DOS Japanese-English translation text file
Others: User dictionary registration
File management
Text file indication
Price: TAKERU version: Y39,800 (incl. taxes)
Package version: Y49,800 (excl. taxes)
Information:
Sakai, Takeru Bureau, Brother Industries Ltd.
Tel: 052-824-2493

Nova introduces PC-Transer/JE

[From *AAMT Journal no.4*, translated by Shravan Vasishth]

1. Background

In 1989, Nova released Transer/ej, a machine system for UNIX workstation which translates from English to Japanese. Its parsing ability, translation speed and user friendly operating environment were greatly appreciated by users. Then, in 1991, Nova adapted the system for IBM DOS/V and NEC PC98 Series environments. With this latest edition, Nova now has a set of four PC-based systems:

* <PC-Transer/ej> English to Japanese translation system designed for the PC and based on the original Transer/ej

* <My Transer/ej> English to Japanese translation system designed for personal as opposed to business use. The system was simplified somewhat so that it requires less memory and is less expensive.

* <PC-Transer/je> Japanese to English translation system (described below)

* <Super Transer/ej> An English to Japanese translation system specially designed to be used with patent-related translations. It has an algorithm which allows the special type of long sentences common in patent documents.

In the past, due to limitations of machine-power, cost, etc., the majority of WS based Nova translation software users were translation professionals, either those in translation companies or in the translation department of major companies. However, the PC-based version is mostly being used by business people, engineers and researchers, etc.

Further, Nova also took steps to develop a version for the Macintosh series of computers (Apple, Inc.) which uses a different environment for text-processing tasks like DTP. The decision to develop a Mac version was largely a result of repeated requests from users of Macs and also due to other factors such as the realization that translation, in a sense, is also a kind of text processing.

On 1 October [1993], Nova simultaneously introduced PC-Transer/ej and PC-Transer/je for the Macintosh. The following is an introduction to PC-Transer/je (Mac version) which is the first Japanese to English translation system for Mac in Japan.

2. The Japanese-English Machine Translation System "PC-Transer/je (Mac version)"

Since its release in December last year [1992], over 900 PC-Transer/ej systems designed for the DOS/V and PC98 environment have been sold to research institutions, the manufacturing industry, etc. The latest Mac version has all the features of this PC version. The details are as follows:

(1) Step by step translation

The biggest advantage of Nova's PC-Transer/je series is its ability to do partial translations of phrases and clauses and to thereby translate the whole sentence compositionally. This feature, developed by Nova, aims to resolve the ambiguity that the syntactic structure of Japanese often engenders.

Sample sentence: *ataeteiru moodo, rinku no nambaa, shoyuusha, baito no saizu oyobi kaku fairu ni taisuru saigo no henkou wo nagai foomatto de risuto shite kudasai.*

First translation: *List the last change for a mode giving, the number of link, an owner, size of byte and each file in a long format.*

The above construction is based on the following interpretation:

```
saigo no henkou wo nagai foomatto de risuto shite kudasai
                    |
                    |---- (ni taisuru)
                             |
                             |---ataeteiru moodo
                             |---rinku no nambaa
                             |---shoyuusha
                             |---baito no saizu
                             |---oyobi kaku fairu
```

However, the correct interpretation of the text is:

```
nagai foomatto de risuto shite kudasai
                    |
                    |--- (wo)
                             |
                             |---ataeteiru moodo
                             |---rinku no nambaa
                             |---shoyuusha
                             |---baito no saizu
                             |---oyobi kaku fairu ni taisuru saigo no henkou
```

To reflect such a structure in the translated sentence, in the step by step translation method, the translation proceeds in two steps:

(1) the clause *kaku fairu ni taisuru saigo no henkou* is translated;

(2) the whole text is translated compositionally on the basis of the results obtained by step (1) above.

These steps are executed by means of simple instructions appearing on the screen. The result is the following, preferred translation:

List a mode giving the number of link, an owner, size of byte and the last change for each file in a long format.

(2) Storage function.

Nova's PC-Transer/je series also has a feature which allows users to store translation patterns easily into a user dictionary. For example, from a sentence such as *watashi wa, kono bunsho wo kikai honyaku suru* one can store the verb *kikai honyaku suru*. Following the pattern "1 ga 2 wo" automatically provided by the system for the verb *suru*, if one stores the verb in the following pattern:

1 translate 2 using a computer

the above sentence in Japanese is translated as: *I translate this document using a computer.*

Several representative patterns are provided and one may choose from among them. The user can provide his own patterns as well. Accordingly, using this function allows the user to configure his own Japanese-English translation system.

(3) Other features.

The system has the following dictionaries: one user dictionary and two technical dictionaries that can optionally used. Fifteen technical dictionaries covering areas such as Computers, Electronic Engineering, Mechanical Engineering, Medical Science, Business, etc., are available.

After the translation is complete, the user can choose the output style: English only, Japanese only, or English and Japanese together. The output file is a standard text file; editing, text formatting, etc. using other DTP software is possible.

We are continuously working on further improving translation quality and the operating environment and are planning to keep upgrading annually.

3. Operating environment, user support and prices

Environment:

OS: Kanji talk 7 or higher
Memory: 8MB or more
HD: 20MB for basic system; average of 8MB for each additional technical dictionary

User support:

Free support via telephone line for installation, general use and translation output questions is available.

Prices:

PC-Transer/ej English to Japanese translation system:	¥ 198,000
PC-Transer/je Japanese to English translation system:	¥ 198,000
PC-Transer/ej + PC-Transer/je set:	¥ 346,000

For further information contact: NOVA, Inc., System Sales Division, #301 Suzusho Bldg., 23 Araki-cho, Shinjuku-ku, Tokyo 160 Japan. Tel: 03-3351-3356; Fax: 03-3351-5766

[All product names appearing in this article are trademarks or trade names of each company.]

LogoVista E to J on Macintosh

[Based on articles in *Language Industry Monitor* no.20 and *AAMT Journal* no.5.]

At the launch of Apple's Power Macintoshes in March, the LogoVista English to Japanese MT system was one of the applications used to demonstrate the capabilities of their new generation of personal computers. LogoVista is the product of a joint venture established by Language Engineering Corporation, based in Belmont, Massachusetts, and the Japanese company Catena. Catena markets its own translation package STAR. However, LogoVista E to J is not an upgrade of STAR, but a quite separate product. In the US the system is sold by LEC, in Japan by Catena. As well as the Macintosh version, the company has developed Unix, Windows (both MS-DOS and DOS/V) and Windows NT versions. The Macintosh version runs on Macintosh II series or Powerbook 140; it requires KanjiTalk 7 or greater, and 8MB or more system memory. Users can choose between "formatted" and "side-by-side" mode. In the formatted mode, the Japanese is displayed in the same format in which the English text was entered. In the other mode, translated

Japanese sentences are displayed on the right side of the English. The system can learn from corrections: if a word or phrase needs to be translated in a certain way in the remainder of a text, the translation is automatically registered in the user's dictionary for use whenever the word or phrase occurs subsequently in the source text, and this version is preserved as an alternative translation for later documents. LogoVista comes with a 100,000 word system dictionary and the company offers twenty two domain-specific dictionaries, as well as facilities for users to compile their own dictionaries.

THAMUS developments

[Extract from *Language International* 6/1, 1994]

The Italian company THAMUS - Consorzio per la Linguistica Computazionale, whose head office is in Salerno but which also has a commercial office in Milan, has decided to venture into the international translation market.

Thamus, a subsidiary of the IRI-Finmeccanica, was set up in 1988 to carry out research into computer science and linguistics for industrial applications in natural language processing. It co-developed the English into Italian and German into Italian language pairs of the LOGOS machine translation technology in Italy. With 28 researchers Thamus is one of the biggest Italian research centres working in the computational linguistics field.

Thamus has an in-house translation group consisting of nine graduate linguists and one computer programmer. This group has gained a great deal of experience working on translations and desktop publishing for big Italian companies such as Siemens-Nixdorf Italia, and Ansaldo and Finmeccanica (IRI). Now Thamus feels it is the right time to venture into the international translation market.

Thamus uses an IBM 4381 mainframe, IBM RS/6000 workstations and Sun workstations. These are interlinked with a LAN, and can be linked with outside companies by means of a special line.

The LOGOS/Thamus machine translation system consists of a dictionary of more than 120,000 words and approximately 35,000 grammar and semantic rules.

Thamus is also working on two advanced translation systems, one from Italian into English, the other from Italian into German. In addition, Thamus is involved in the EUROLANG project which, over a period of three years aims to develop a second generation machine translation system centred around English and the major European languages.

News from Systran

[Extracts from Systran's newsletter *Translation News* vol.3 no.1, Winter 1993/94.]

Continuation Contract Awarded for C Conversion. The US Air Force has awarded a continuation of last year's software conversion contract. With the C software versions, translations will operate on desktop computers, such as a 486 PC or Unix workstation. The C-conversion contract benefits both the government and Systran. By funding development, the government can operate the new C-based software on nay of its estimated millions of government-based desktop computers and UNIX workstations. At the same time, Systran can continue to offer enhanced versions of its software for the benefit of its expanding customer base.

Systran and Berlitz Form Joint Relationship. Systran Translation Systems Inc. and Berlitz Translation Services Inc. have announced the formation of an alliance. Systran has begun offering the well-known Berlitz name as an option to Systran customers for post-editing services. Systran's in-house machine translation services uses a post-editing process, to ensure absolute accuracy. Prior to the recent explosive growth of MT services, Systran subcontracted to smaller translation agencies.

Systran Signs Largest Commercial Contract. Following the introduction of the Machine Translation Services Department, the largest commercial services contract in Systran's history was recently awarded to Systran and Berlitz Translation Services, Inc. The contract calls for translating CAD/CAM user manuals and other technical documentation, and localizing software. Machine translations will be made from English into French, German, Italian and Spanish.

Translation Sales Jump Forward. For the second consecutive year, Systran's translation services sales have risen nearly 40 percent from those of the previous year. The marked increase over the past two years is derived from the manufacturing sector's growing awareness of Systran as a complete service company. New customers are often surprised at the cost efficiency. Typical charges include \$0.05 per word for the MT processing with an additional \$0.08 to \$0.10 cost for the post-translation editing, language depending. Where format integrity is important, such as with technical documentation, machine translators can expect savings over traditional services of thirty to forty percent.

Mainframe Computers for MT - The End of an Era. Systran will soon begin to convert its entire operation from the company's mainframe computer to independently linked desktop platforms. For the past 25 years, Systran linguists and translators have depended upon the IBM 370 series mainframe computer for all software development and maintenance. By March of 1994, most linguistic development will be carried out on an IBM model PS/2 translation workstation. The workstation is equipped with an IBM p/370/A emulator board and the OS/2 v.2.1 operating system.The PS/2 will act as server to a network of high powered 486 PC terminals.After a year of vigorous product testing, Systran has signed on with IBM as an OEM distributor for the PS/370/A emulator board.

User Access to Dictionaries due January '94. Until recently, the only means by which a user had to augment the vast bulk of vocabulary data stored on Systran's mainframe computer was through a supplementary stand-alone dictionary. This dictionary has been an integral feature of the Systran Express program. Systran Express grants user access to the mainframe translator by way of a PC terminal and a modem. Such custom dictionaries are limited to single words and noun phrases. These naturally add to the translation processing time.

Systran programmers have now developed faster ways to speed their own customizing task. Containing over two million entries, Systran dictionaries contain the largest machine-based vocabularies in the world. Needed was a user-friendly software manager to help programmers sort and augment the vast and complex database. The tool has been named the "Dictionary Builder".

The Dictionary Builder uses a graphical user interface or GUI. With the system soon to be operating on desktop platforms, users will now have informative icons and a mouse at their fingertips to help create faster, more linguistically accurate and complex dictionaries. Assembly line coding is also planned to permit faster terminology and expression coding. An entry rate of 5,000 entries per month is anticipated.

With the new Dictionary Builder, Systran customers may also develop highly specialized glossaries specific to their industry and line of products.

Globalink Offers New Software

[From *Language International* 6/1/, 1994]

Globalink, Inc. adds to its line of software for foreign language translation with a product that translates English documents into Russian.

The product POWER TRANSLATOR *PROFESSIONAL*™ (version 3.0 for DOS), complements the company's Russian to English language translation software, allowing Globalink to offer bi-directional computer assisted translation for English to/from Russian.

The software, it is claimed, translates full sentences at more than 20,000 words an hour with idiomatic accuracy of more than 90%. It has a 250,000+ word and phrase general dictionary. In addition, users can purchase a business subject dictionary with words and phrases specific to the business profession. The subject dictionary allows the user to obtain an accurate translation of terms such as "bull market" without having to rely on the translation of the words "bull" and "market" from the general dictionary. Subject dictionaries are available from Globalink, or users can create their own.

POWER TRANSLATOR *PROFESSIONAL*™ can be used with Perceive/Globalink Edition™ optical character recognition (OCR) software, which recognises Russian Cyrillic. The software accepts ASCII or .TXT files. It has a number of features including multiple translations based on part of speech, linking and display of related terms, variable translation of prepositions, automatic inflection of semantic units, automatic morphological inflection, automatic verb conjugation and noun pluralisation, alphabetical browsing of dictionaries, and more.

The software will retail for $1,450 and is available through distributors or by calling Globalink directly. Users must have VGA systems with 256K of VGA memory.

POWER TRANSLATOR *PROFESSIONAL* is also available in Spanish, French or German to/from English for DOS, OS/2, UNIX and Macintosh platforms. The company's POWER TRANSLATOR™, which recently won Discover magazine's 1993 Award for Technological Innovation, is available in Spanish, French or German to/from English for DOS platforms. It is designed to meet the needs of personal computer users, students, and businesses with general translation requirements.

EUROLANG - The Translator-92s Workbench

[From information leaflet]

Sonovision ITEP Technologies (SITE), a division of the CORA Group, has announced a suite of Natural Language Translation tools under the name EUROLANG OPTIMIZER.

EUROLANG OPTIMIZER is sophisticated, easy to understand and simple to use. It is fully integrated with Microsoft WORD 6.0. The user simply accesses all EUROLANG functions from within MS-WORD by means of a Custom Toolbar.

The EUROLANG implementation includes Translation Memory and an integrated TermBank. EUROLANG reduces translators-92 workloads by searching the Translation Memory Server for Perfect and Fuzzy (approximate) Matches for previously translated sentences; and the TermBank for known Technical Terms. The document is accordingly marked up by the server and sent to the Client station, where Perfect Matches, Fuzzy Matches, and recognized Technical Terms are each highlighted on the screen in specific colours to permit speedy editing and creation of the output document. Any unrecognized sentences and technical terms are then translated by a human or EUROLANG-92s Machine Translation option and by means of document administration, are added to the Translation Memory and Termbank, respectively. This process continually enhances the system-92s ability to find matches in subsequent Pre-Translation runs.

This Client-Server Architechture has shown reductions in translation workloads of up to 42% in repetitive texts. EUROLANG learns as you use it. The more you use the system, the better it gets.

Another key feature - EUROLANG includes alignment tools so that you can very easily build your own translation memory from existing previously translated documents and a complete range of administration tools for the most complex logistics.

EUROLANG has been chosen by Microsoft and other leading American corporations.

EUROLANG OPTIMIZER is available on:

1. PC Platform - Windows NT/MS-SQL (Server) and WINWORD 6.0 (Client).
2. UNIX- Solaris X11/Motif, Sun OS 4.1.3 - ORACLE or SYBASE, FRAME or INTERLEAF

Source Languages	Target Languages		
Q2/94	Package 1	Package 2	Package 3
	Q2/94	Q3/94	Q4/94
English	Dutch	Czech	Chinese
German	English	Danish	Japanese
French	French	Finnish	Romanian
Spanish	German	Greek	Russian
Italian	Italian	Hungarian	Turkish
	Portuguese		
	Spanish		
	Swedish		

For more information, contact: Lutz Graunitz (800) 565-5650 or (416) 496-8510, e-mail: lutz@sni.ca

PROJECTS, DATABASES, JOURNALS

Machine Translation Research in Indonesia

Darmawan Sukmadjaja

[From AAMT Journal no.4, translated by Shravan Vasishth]

Indonesian, or Bahasa Indonesia, belongs to the Malayo-Polynesian or Austronesian family of languages. The Austronesian languages are generally sub-grouped into Indonesian, Melanesian, Polynesian and Micronesian. On August 17, 1945, the day of independence, Bahasa Indonesia was proclaimed as the national language of the new republic and over the last few decades, it has become more and more popular. Nowadays it is the sole language used as the vehicle of instruction, from elementary school through to university.

The morphological process of affixation in Bahasa Indonesian (BI) marks it out as distinct from other languages. A discussion of affixation always occupies a position of importance in the grammar of BI, since this is the main word-forming mechanism in the language. Although until now there has been a tendency to limit research in affixation to investigating its appearance in conventional usage, this characteristic feature of BI makes it an attractive subject for study in the area of NLP. It is worth noting that so far no comprehensive research has been done on BI in this field.

BPPT, a government agency responsible for the assessment and application of new technology, has become aware of the importance of NLP, especially in MT. The development of our country depends on the absorption and transfer of technology from developed countries. However, language presents a major barrier to this process. We believe that MT can solve this problem.

We also believe that we cannot fully depend on other countries for this technology, since native speakers of BI will be needed in order to obtain satisfactory results. In view of this situation, BPPT initiated research in NLP and MT in 1986. Our young researchers succeeded in developing a prototype of an English-Indonesian MT system, called EICATS (English-Indonesian Computer Aided Translation System).

In 1987, as a representative of the Indonesian government, BPPT joined the Center of International Cooperation for Computerization (CICC) project for R&D of the Multilingual Machine Translation System between Japanese, Chinese, Indonesian, Malay and Thai. This project has accelerated the pace of research activity in Indonesia. Several research institutions and universities are participating in this project. The National Language Center and Atmajaya University are responsible for the linguistic aspects. A large number of students are interested in doing research in this area. We believe they will become energetic researchers in the near future.

Seven years have gone by since the CICC project was announced. During the six years of Indonesia's involvement with CICC, we have been able to study and research many things concerning MT-related technology, linguistics and software engineering. An Indonesian Electronic Dictionary, indispensable for language processing, has been developed. INAS, Indonesian Analysis System, has also been developed for analyzing BI. We are currently processing a set of corpora in order to obtain statistical information on Indonesian word order and frequency of occurrence. We are also interested in R&D in user-interface for MT systems, such as editors, translation support systems, maintenance systems, and distributed environments for MT systems. The MT support systems that we are currently developing are: User Operation Panel (UOP), Bilingual Editor, and I/O systems. When ready, these will serve both as important tools for making the use of MT more widespread and as an aid for human translators. Many problems need to be solved in this area. Close cooperation between the fields of computers and linguistics is required. We believe that the increase in MT researchers in the near future and the experience that we have gained from the CICC project will be key factors in overcoming remaining problems.

Information from: Darmawan Sukmadjaja, Directorate of Electronics and Informatics Technology, Agency for the Assessment and Application of Technology (BPPT), Indonesia. Email:bppt!daruma@go.id

TSNLP
Test Suites for Natural Language Processing

Siety Meijer, Lorna Balkan

The University of Essex, UK, in conjunction with Aérospatiale, France, Deutsches Forschungszentrum für Künstliche Intelligenz GmbH. (DFKI), Saarbrücken, Germany, and Istituto per gli Studii Semantici e Cognitivi (ISSCO), Geneva, Switzerland, has just embarked upon a joint research project on the design and use of test suites in Natural Language Processing. The official start of the project was on 1 December 1993. The project is part of the CEC's Linguistic Research and Engineering (LRE) programme.

The project aims to produce a set of guidelines for the construction of test suites for a range of NL products, concentrating on grammar checkers, parsers and controlled language checkers. Important issues in the design of test suites will be identified and addressed for each application type. In addition, the project aims at producing test data in three languages: English, French and German.

The project will also develop a number of tools to facilitate the construction and use of test suites. An automatic generation tool based on a simple grammar will be developed as a first step to more flexible test suite construction. The test suite will also be mounted on a database, which will allow easy access and manipulation of the data.

The results of the project will become public domain, and will, hopefully, provide impetus for a more general development of test suites and other evaluation tools.

The first stage of this project involves reviewing what test suites and guidelines for test-suite construction exist already. Public domain test suites are in short supply. We would like to hear from anybody who has been involved in either the construction or design of test-suites in however small a capacity, and if possible, have access to their test-suites however unsystematic or partial. We will treat information we are given with any degree of confidentially that is desired.

Contact: Ms Siety Meijer, or: Ms Lorna Balkan, Department of Language and Linguistics, University of Essex, Wivenhoe Park, Colchester, Essex CO4 3SQ, U.K. (Fax: +44 206 872085; Tel: +44 206 872091/92; Email: meyes@essex.ac.uk, balka@essex.ac.uk)

The Linguistic Data Consortium

[Since some MT researchers may not know about this programme the following extracts have been reproduced from the Linguist list.]

There is increasing interest in computer-based linguistic technologies, including speech recognition and understanding, optical and pen-based character recognition, text retrieval and understanding, and machine translation. In each area, we have useful present-day systems and realistic expectations of progress.

However, because human language is so complex and information-rich, computer programs for processing it must be fed enormous amounts of varied linguistic data - speech, text, lexicons, and grammars - to be robust and effective. Not even the largest companies can easily afford enough of this data to satisfy their research and development needs. Researchers at smaller companies and in universities risk being frozen out of the process almost entirely.

Therefore, the Defense Advanced Research Projects Agency (DARPA) has formed a broadly-based consortium that will include companies, universities, and government agencies. Known as the Linguistic Data Consortium (LDC), this organization will make it possible to share pre-competitive development costs widely. An initial grant of $5 million from DARPA will amplify the effect of contributions from a broad membership base, so that there is guaranteed to be far more data than any member can afford individually. Additional government investment in the technologies supported by the LDC is likely in the future.

The operations of the LDC will be closely tied to the evolving needs of the research and development community that it supports. Research opportunities will increasingly depend on access to the consortium's materials, and membership fees will be set at affordable levels. At the same time, a significant fraction of the consortium's budget will come from its industrial members.

Parallel efforts are underway in Europe and Japan. Productive relationships will be negotiated with these efforts, on the principle of open access across continental boundaries to the raw materials of technological progress.

Most of the LDC's work will be done under contract by members of the community of researchers that it serves. Organizational and administrative functions will be provided by a director and staff, located at the University of Pennsylvania, which has been selected as the LDC Host Institution. Oversight is provided by the LDC Board, which consists of the director, up to three individuals appointed by DARPA to represent the community of researchers, and one representative appointed by each LDC Senior Member.

The LDC aims to ensure that no researchers are excluded by virtue of genuine inability to pay, while at the same time raising sufficient funds through membership fees to support on-going work. Therefore the LDC Board has set yearly membership fee at $2,000 in the case of not-for-profit entities, and $20,000 in the case of for-profit entities. Cases in which these fees pose insuperable obstacles will be considered and resolved on an individual basis by the Board.

In addition to its contribution of money, DARPA has contributed the following speech and text corpora to the LDC:

* the TIMIT speech corpus;
* the Resource Management (RM) speech corpus;
* the Air Travel Information System (ATIS) speech corpus;
* the Penn Treebank annotated text corpus;

* the MUC corpus of FBIS terrorist reports;
* the SWITCHBOARD speech corpus;
* the Air Traffic Control (ATC) speech corpus;
* the TIPSTER/TREC text corpus.

The need for the LDC arises out of on-going work in areas such as speech recognition, machine translation, and full-text information retrieval. A crucial judgment about the value of its products will be provided by their immediate application to research projects in these fields. As the affected technologies evolve, and as new linguistic technologies emerge, the needs for data will change, and the activities of the LDC will track these changes closely.

To join the LDC, or to get more information about its operations, please contact: Elizabeth Hodas, 441 Williams Hall, University of Pennsylvania, Philadelphia, PA 19104-6305 (Tel: +1 (215) 898-0464; Fax: +1 (215) 573-2175; Email: ehodas@walnut.ling.upenn.edu)

Flash Information

ISIR, the Integrated Service of Information Resources, of the Centre for Information Technology Innovation (Industry Canada) disseminates a selective bibliography and information briefs aimed at R&D managers, researchers, and professionals within the field of information technologies (computers and computing; software engineering; natural language processing; multimedia systems; information storage; interchange and retrieval; work organization; etc.).

This weekly publication, Flash Information, is distributed via e-mail and is free until further notice. To receive the publication send your e-mail address to flash@citi.doc.ca.

Journal of Natural Language Engineering

[Publicity announcement from the Linguist list 5-268]

"Natural Language Engineering" is a new journal, to be launched by Cambridge University Press in 1995. It is an international forum for the dissemination of results concerning the theory and practice of applied natural language processing. The focus is on publications addressing research and development issues fundamental to engineering of natural language technologies. A principal concern will be systems which can operate in realistic application environments, in terms of their scale, robustness, feasibility, maintainability, usability and system integration.

The editors are particularly anxious to respond to identified needs in the research community. The field currently lacks an outlet dedicated to communicating technology-oriented work. There is also a lack of easily accessible reference materials discussing the impact of research in computational linguistics, computer science, artificial intelligence, and cognitive psychology on the task of engineering for practical natural language processing applications.

Original articles are sought, addressing issues in all areas of linguistic engineering; these include, but are not limited, to: adaptive systems; corpus processing; delivery of assistance (e.g. help systems, explanation); dialogue management; front ends to computational systems, both single modality and multimodal; grammar and style checking; information management and access; language teaching aids; lexicon acquisition and implementation; linguistic knowledge bases (e.g. lexica, grammars, term banks); localising for multilingual systems; machine translation; multimedia authoring and delivery environments; performance evaluation; speech and natural language integration; text analysis and content extraction; text generation; tools for natural language processing. Articles may discuss separate technologies, complete applications, system evaluations, and limitations of particular designs.

Executive editors: Roberto Garigliano (University of Durham); joint editors: John Tait (University of Sunderland), and Branimir Boguraev (Apple Computer, Inc.).

Natural Language Engineering is published four times a year in March, June, September and December. Four issues form a volume. Volume 1 will be published in 1995.

Authors from Europe, Asia and Africa should submit papers to: Roberto Garigliano, Laboratory for Natural Language Engineering, Computer Science Department, University of Durham, South Road, Durham DH1 3LE, United Kingdom (tel: +44 91 374 2639; fax: +44 91 374 2560; email: Roberto.Garigliano@durham.ac.uk)

Authors from North and South America and Oceania should submit papers to: Branimir Boguraev, Advanced Technologies Group, Apple Computer, Inc., One Infinite Loop, MS: 301-3S, Cupertino, California 95014, USA (tel: +1 408 974 1048; fax: +1 408 974 8414; email: bkb@apple.com)

Electronic submissions are strongly encouraged. For guidelines for authors please contact as above.

Machine Translation

Special Issue on Building Lexicons for Machine Translation

Editor: Sergei Nirenburg
Guest Editors: Bonnie J. Dorr and Judith L. Klavans

[Call for papers announced in Linguist list]

The journal of *Machine Translation* is planning a Special Issue on the Lexicon in Machine Translation (MT). The lexicon plays a central role in any MT system, regardless of the theoretical foundations upon which the system is based. However, it is only recently that MT researchers have begun to focus more specifically on issues that concern the lexicon, e.g., the automatic construction of cross-linguistically valid lexical-semantic and knowledge-based representations for use by multi-lingual systems. The need for large dictionaries is overwhelming in any natural language application, but the problem is especially difficult for MT because of cross-linguistic divergences and mismatches that arise from the perspective of the lexicon. Furthermore, scaling up dictionaries is an essential requirement for MT that can no longer be dismissed; researchers need to move from toy-dictionary MT systems into larger-scale MT systems so that they will be in a better position to demonstrate the validity of the theoretical underpinnings of their systems.

The intent of this Issue is to address critical issues concerning the automatic and semi-automatic acquisition of lexical representations for MT dictionaries. Among traditional approaches to constructing dictionaries for natural language applications has been the massaging of on-line dictionaries that are primarily intended for human consumption. Given that many natural language applications have focused primarily on syntactic information that can be extracted from the lexicon, these methods have constituted a reasonable first-pass approach to the problem. However, it is now widely accepted that natural language processing in general, and MT in particular, requires language-independent conceptual information in order to successfully process a wide range of phenomena in more than one language. Thus, the task of lexicon construction has become a much more difficult problem as researchers endeavor to extend the concept base to support more phenomena and additional languages. Added to this is the standard size, coverage, efficiency trade-off, combined with the fundamental question of anticipated vs actual functionality.

High-quality original research papers are invited on issues relevant to this topic including, but not limited to:

- Lexical levels required by a machine translation (syntactic, lexical semantic, ontological, etc.) and interdependencies between these levels.
- Automatic procedures for the construction of lexical representations.
- Semi-automatic methods for the acquisition of lexical knowledge.
- Use of existing resources and aids for transforming these resources into appropriate representations for MT.
- Augmentation of statistically driven corpus analysis with linguistically motivated techniques for extracting lexical knowledge.

- Role of bilingual dictionaries, including example sentences and phrases. Extraction of information from pairwise data in dictionaries.

- MT mappings (transfer, interlingual, statistically based, memory-based, etc.) and the effect of these mappings on the representation that is used in the lexicon.

- Language universals in the lexicon and the construction of an interlingua for MT.

- Incorporation of lexical/non-lexical knowledge for selection of suitable candidates for target constructions in MT.

- Accommodation of MT divergences and mismatches in the lexicon; implication for automatic construction of lexicons.

Deadline for submission of articles: July 15, 1994. Articles may be submitted in hard-copy, electronic (either plain text or .ps format) to either guest editor. If submitting hard-copy, please send four copies of the paper.

Bonnie J. Dorr, Department of Computer Science, A.V.Williams Building, University of Maryland, College Park, MD 20742 (Email: bonnie@umiacs.umd.edu; Fax: 301-314-9658)
Judith L. Klavans, Department of Computer Science, Mudd Building Room 420, 520 W. 120th Street, New York, New York 10027 (Email: klavans@cs.columbia.edu; Fax: 914-478-1802)

Consortium for Lexical Research Newsletter

From the Computing Research Laboratory
New Mexico State University
Edited by: Jim Cowie and Katherine Mitchell

In February the Computing Research Laboratory mailed the Consortium for Lexical Research Newsletter No. 11 to a large list of researchers in linguistics and computational linguistics. Anyone wishing to be put on their mailing list in order to receive the CLR newsletter every two months should send a request to: lexical@crl.nmsu.edu.

Newsletter no.11 (February 28, 1994) discusses machine readable dictionaries made available through CLR, and describes some recently acquired parsers. The next newsletter will describe wordlists stored in the CLR archives, and a new service to CLR members, called Resources, which will centralize information on ftp sites, organizations, projects, publications, etc. of interest to the natural language processing research community.

Consortium for Lexical Research

Katherine A. Mitchell

The Consortium for Lexical Research is designed to serve as a repository for software and resources of importance to the natural language processing research community. Sharable resources, and the task of centralizing lexical data and tools, are of foremost concern in lexical research and computational linguistics. It is our objective to help alleviate the repeated re-creation of basic software tools, and to assist in making essential data sources more generally available.

CLR maintains a public ftp site, and a separate library of materials only for members of CLR. Currently CLR has about 60 members, mostly academic institutions, and almost every major natural language processing center in the U.S. belongs. Access to the members only materials is strictly regulated by password and userid.

Perhaps the best way to become familiar with CLR is to look at our catalog of current holdings. This is available by using ftp anonymous to clr.nmsu.edu (128.123.1.12). The file you need to 'get' is "catalog.ps" for a postscript version, or "catalog" for a simple ascii version.

We have three categories of membership; all fees are annual:

1) US government agencies ($500.00)

2) research organizations, including universities, for which the annual fee is $250.00.

3) commercial organizations for which the annual fee is $2500.

In addition, a member must sign the membership agreement which specifies distribution limitations by the member. A member then gets access to the members-only archives, and (depending on approval of the provider) the ability to access the heavily encumbered items. Heavily encumbered items include machine-readable dictionaries; however, publishers charge additional fees for these items.

A member may join the Consortium and provide materials in exchange for membership. If your company or organization has data or software that they are willing to deposit in the CLR, membership fees can be waived partially or completely. I am very interested in this exchange; CLR is making a very concerted acquisitions effort in the next 6 months.

If you wish to receive the membership agreement, please let us know.

Consortium for Lexical Research, Computer Research Laboratory, New Mexico Sate University, Box 30001/ 3CRL, New Mexico State University, Las Cruces, New Mexico 88003 (Email: lexical@nmsu.edu; tel: (505)-646-5466)

CELEX database

Since 1986, the Dutch national Expertise Centre CELEX (Centre for Lexical Information) has been constructing large electronic databases containing various types of lexical data on present-day Dutch, English and German. CELEX makes this information available to institutes and companies engaged in language and speech research and in the development of language and speech oriented technological systems. Using the specially-developed program FLEX, you can access the databases with ease -- no technical expertise is necessary -- and extract information which matches the detailed requirements you have. In addition, CELEX can offer assistance with respect to related research and development projects.

If you are interested in finding out more about CELEX, then please get in touch with us. We can send you copies of our introductory booklet, plus back issues of the five newsletters so far published, and answer any specific questions you might have. In many cases a `trial' account can be set up to let you look round the databases before making any financial commitment

You can send email: CELEX@CELEX.KUN.NL (Internet), CELEX@HNYMPI52 (EARN/BITNET), or write to the following address: CELEX -- Centre for Lexical Information, University of Nijmegen, Wundtlaan 1, 6525 XD Nijmegen, The Netherlands

ASSOCIATION NEWS

AMTA Conference Gets into High Gear

Preparations are now in full swing for AMTA's first-ever conference, "Technology Partnerships for Crossing the Language Barrier," to be held in Columbia, Maryland, on 5-8 October 1994. The goal of CAMTA-94 is twofold: above all, to share the latest information on every front in the machine translation field, and, at the same time, to encourage all who are interested in language technology - actual and potential users, vendors, developers, and researchers - to focus on ways in which they can team up to improve communication across language barriers. Three days of sessions, some of them concurrent to ensure something for everyone, will be preceded by tutorials on the afternoon of 5 October on "Choosing an MT System" and "Intellectual Property Rights in MT." A welcome reception will follow in the exhibit area.

CAMTA-94 promises to be both a productive and a memorable event. AMTA members have had a voice in deciding on the program and the venue of the conference. A poll conducted last

January disclosed the topics in which they were most interested as well as their preferences as to date and site.

The topics of greatest interest to AMTA members, starting from the top, are: knowledge-based MT, future directions, the concept of translatability, MT on PCs, MT evaluation, current solutions, translator workstations, and "workstation" MT. Program co-chairs Eduard Hovy and Joseph Pentheroudakis are now hard at work weaving these topics into a series of panels, papers, demonstrations, special interest groups, and other events for everyone's enjoyment.

The majority of AMTA members preferred to meet during the month of October (many wanted it immediately before of after the annual conference of the American Translators Association) at a site in "scenic Maryland." The arrangements that have been made will definitely fill the bill. While forging partnerships, participants will also be making memories in the lovely woodland setting of Columbia Inn on the shore of Lake Kittamaqundi in Columbia, Maryland. In the first week of October the weather will still be mild, but cool evenings should have already brought a hint of fall color. Situated on 10 wooded acres, Columbia Inn connects with footpaths around the Lake and offers ample opportunities for recreation. At the same time, it is within walking distance of first-class restaurants, entertainment, and 200 shops in Columbia Mall (including 35 eateries). Exciting excursions are being planned. Inexpensive transportation can be arranged to nearby Washington, Baltimore, or Annapolis, and evening tours to any of the three cities are available for $5.00. Columbia is an easy 15-minute ride from Baltimore-Washington International Airport.

The registration fees (in US dollars) are as follows:

	Until 1 Aug	After 1 Aug
Conference registration:		
Active member of AMTA	125	150
Associate member of AMTA	175	200
Member of AAMT or EAMT	175	200
Nonmember	195	220
Tutorial fee:		
Active or corp/inst member of AMTA	75	95
All others	100	120

The early registration rate will be guaranteed for those who return the form at the end of this issue of MTNI.

The CAMTA-94 meeting rooms and exhibit area directly face the Lake, as do many of the guest rooms in the new higher-rise "Tower" (early reservations recommended for the most commanding views and nonsmoking floors). On the other hand, participants may prefer the newly renovated but still cozy "Inn" section, truly nestled in the woods, with openable windows. Covered parking is available at no cost. The daily rate for either sections, single or double occupancy, is $79.00 plus tax; suites are $225.00. These are special AMTA conference rates guaranteed to participants through 20 September; participants should make reservations directly by calling (800) 638-2817 or (410) 730-3900, specifying "AMTA" when they call.

Further information on CAMTA-94 will be found in the 'Conference Announcements' section later in this issue.

MT Yellow Book Goes International

The MT Yellow Book, published by AMTA last July, will now be expanded to include members of all the IAMT regional associations. Internationalization of the Yellow Book is an important step toward the the increased exchange of information about MT.

In essence, the Yellow Book is a directory of individuals, institutions, and companies that are involved in, or interested in, machine translation. The first edition, initiated by AMTA and largely focused on the Americas, has already met an important need. Until its appearance, there was no centralized way of facilitating communication among the people in the MT field. The 1995 edition will meet this same need at the world level.

Like the current MT Yellow Book, the 1995 edition will have three sections: the first, a master directory of all names, with an indication of regional affiliation and cross-reference to all classified listings; the second, full listings for all members of each of the regional associations - AAMT, AMTA, and EAMT - by region (those who don't want to be included may so indicate by checking that box on the form); and the third, a section of classified listings. This layout can be examined in the current edition, copies of which can still be ordered (see below.)

Those who want to be listed in the upcoming international edition should fill out the form at the end of this issue of MTNI and return it to their regional secretariat. Active members of AAMT, AMTA, and EAMT are entitled to one free classified listing. Other conditions for classified listings are explained on the form.

AMTA is still selling the first edition of the Yellow Book, a handy locator of AMTA members and others who are involved in, or interested in, machine translation. Copies are available from AMTA Headquarters and may be requested using the form at the end of this issue. The price is $6.00 plus $4.00 for postage and handling.

AAMT meetings: August 1993 to March 1994

The *AAMT Journal* nos. 5 and 6 (December 1993 and March 1994) records the following meetings of the Asia-Pacific Association for Machine Translation.

Board meeting:

10 September 1993: approval of new office address; financial report for the period of April/August

23 March 1994: business and financial planning for fy 1994.

Management Committee:

26 August 1993: nomination of new office address; financial report for the period of April/July; activity plans for the latter half of fy 1993.

21 September 1993: new office address and rent contract approved; sanction of "Translation Fair" (JTA) denied, of "Translation Day" approved.

19 October 1993: work gropus activities; plans for Workshop and Get-together for MT Users; Market Forecast Work Group inaugurated.

17 November 1993: expenses for office removal reported; reports of the IAMT 2nd general assembly and its resolutions; subsidies for Coling '94; activities outline for fy 1994.

14 December 1993: Work Groups activities reported; outline for MT Users' Workshop.

25 January 1994: financial report for September/December; report on IAMT Steering Meeting; work group activities.

22 February 1994, reports on sanctioned events; priority projects for fy 1994

11 March 1994, budget outline for fy 1994; proposed co-sponsoring events; expected balance sheet for fy 1994

Treasury Committee:

20 August 1993: expenses for office removal expected; rent contract for new office examined.

Joint Editorial Committee:

20 December 1993: assistance to IAMT newsletter.

Systems Utilization Technic Work Group:

14 September 1993: activities plans; project goals and project managers nominated.

8 October 1993: project plans; sharing user dictionaries; MT users' utility survey planned; details for MT Users' Workshop and MT Users' Get-together.

18 November 1993: invited lecture (Mr.T.Sagara); sharing user dictionaries; details for MT users' utility survey; teaching material for MT Users' Workshop; job protocol for MT users.

21 January 1994: plan for MT users survey; dictionary sharing; total job flow for MT and translation; hearing from professional MT user

23 February 1994: details for the user survey

23 March 1994, User survey: Nagase & Co.Ltd.

24 March 1994, User survey: IBS Co., Ltd.

System Evaluation Work Group:

21 October 1993: reviews of AAMT Survey Report; direction of activities shifted for language-oriented; source texts not suited for MT.

19 December 1993: invited lecture (Mr.T.Ashizaki); MT bibliography (Mr.S.Kameda); MT unsuited source texts.

28 January 1994: invited speaker from JEIDA; assignment to extract & process original texts unsuited to MT and rewriting the input.

9 March 1994: reports & post-editing the assignment.

Market Forecast Work Group:

28 October 1993: needs for market forecast; reference materials and survey reports review; outlook of translation market.

16 November 1993: references and survey reports; study plans; keywords; non-MT translation market.

16 December 1993: references and study plans.

27 January 1994: references for survey; model for market forecast.

1 March 1994: structural change in MT market; potential demands and PCC MT market.

Sanctioned Events:

26-27 August 1993: Technical Writing Symposium in Tokyo.

30 September 1993: Translation Day events and lectures by the Japan Translation Association in Tokyo.

2 October 1993: exhibition and lectures by the Japan Translation Federation in Tokyo.

4 February 1994: Translation Fair '94 in Tokyo, organized by Japan Translation Association. A seminar, lectures and panel discussion. MT demonstration.

READERS' FORUM

This section of MT News International is for readers to express personal views of issue in the field of MT. The inclusion of an item does not, of course, imply the endorsement of the views expressed by either the editors or the International Association for Machine Translation.

With this issue we include the first of regular contributions from Colin Brace, editor of Language Industry Monitor.

Notes from the Field

Colin Brace

An Obsolete Distinction

Since time immemorial, the MT world has been accustomed to differentiate between Machine- or Computer-Aided Translation and Machine Translation proper. The two group's respective proponents and practitioners largely worked in isolation from each other, each dedicated to its own pursuits. Traditionally, MT researchers and developers developed huge software edifices on the basis of complex linguistic models (i.e., METAL and Logos) or massive lexicons (Systran). The CAT world, meanwhile, has expended most of its efforts modeling the way translators work and has tried to provide electronic equivalents of traditional tools that rely on a minimum of linguistic processing.

Recent developments make it clear, however, that these worlds are coming together, and the distinction we have maintained between MT and CAT is no longer relevant -- or useful. With its translation workstation TM/2, IBM places the translator rather than the translation (or the translation system) squarely in the middle of the equation. While IBM has not released its MT system LMT as a product yet, it has demonstrated the integration of LMT within TM/2 as an optional service. The Eurolang consortium, which has just released the first version of its translation workstation package, further consolidates this trend. With Optimizer, Eurolang not only heralds a powerful new generation of word- or document-processor based translation tools but also brings MT closer to the desktop. Going further than IBM, Eurolang promises to deliver the first release of its MT system, the successor to METAL, later this year. It will plug into the Eurolang architecture, and will be offered, as with TM/2, as one service among an array of many at the disposal of the translator, functioning in conjunction with the termbase and the translation memory, and, if need be, distributed across a network.

With the move towards increased interaction and other developments on the horizon, the machine translation landscape is unequivocally changing, and all for the better. For today's MT developers, the moral of the story is clear: ignore the desktop at your peril!

FROM THE ARCHIVES...

The Georgetown-IBM demonstration, 7th January 1954

John Hutchins

On the 8th January 1954, the *New York Times* carried a front-page report of the demonstration on the previous day at the IBM headquarters in New York of a system which translated some Russian sentences into English. It was the result of a joint project by IBM staff and members of the MT team recently gathered at Georgetown University under the leadership of Leon Dostert.

The impact of this demonstration was profound, possibly the most influential publicity that MT has ever received. Certainly, it is one of the few times that MT has been front-page news and for this reason it makes interesting reading forty years later, and the full text of the report is reproduced.

<< **Russian is turned into English by a fast electronic translator** by Robert K.Plumb

A public demonstration of what is believed to be the first successful use of a machine to translate meaningful texts from one language to another took place here yesterday afternoon.

This may be the cumulation of centuries of search by scholars for "a mechanical translator." So far the system has a vocabulary of only 250 words. But there are no foreseeable limits to the number of words that the device can store or the number of languages it can be directed to translate.

Scholars and scientists who worked on it believe that within a few years the system may greatly increase communication, particularly in technical subjects, by making translation quick, accurate and easy.

The demonstration was at the headquarters of the International Business Machines Corporation, 590 Madison Avenue. It is the result of cooperative research by scientists of the corporation and scholars of the Georgetown University Institute of Languages and Linguistics in Washington.

The "mechanical" part of the translation system, which is mostly electronic, is a standard commercial model of the largest International Business Machines "stock" computer. This device, called the IBM Type 701 Electronic Data Processing Machine, was put on the market last April. Since then twelve of the machines have been sold to commercial, military and university computation laboratories.

The "literary" part of the system is a mechanical model of language devised at Georgetown by Prof.Leon Dostert and Dr.Paul Garvin. The corporation's share in the project was conducted by Dr.Cuthbert C.Hurd, director of its Division of Applied Science.

In the demonstration, a girl operator typed out on a keyboard the following Russian text in English characters: "Mi pyeryedayem mislyi posryedstvom ryechi" The machine printed a translation almost simultaneously: "We transmit thoughts by means of speech." The operator did not know Russian. Again she types out the meaningless (to her) Russian words: "Vyelyichyina ugla opryedyelyayatsya otnoshyenyiyem dlyini dugi k radyiusu." And the machine translated it as: "Magnitude of angle is determined by the relation of length of arc to radius."

Several short messages, within the 250-word range of the device, were tried. Included were brief statements in Russian about politics, law, mathematics, chemistry, metallurgy, communications and military affairs. The sentences were turned into good English without human intervention.

The heart of the system is the mechanical model of language devised at Georgetown. There the scholars first assembled a 250-word vocabulary in Russian covering the seven broad fields. Then they determined the rules of syntax required for a meaningful statement and reduced them to six instructions for the data-processing calculator.

These instructions are introduced into the calculator's short-term electrostatic "memory" with punch cards. The cards tell the machine how to cope with syntax.

In translating, for instance, a word "A" which precedes a word "B" in Russian, may be reversed in some cases in English. Each of the 250 words is coded for this inversion. Sometimes words must be inserted in the English text, sometimes they must be omitted, following code instructions.

When there are several possible English meanings for a Russian word, the instructions tell the machine to pick out the meaning that best fits the context.

Foreign words are typed on a keyboard that punches I.B.M. cards. These are fed into the calculator, where they encounter the vocabulary, also punched on cards. On a standard printer meaningful English texts emerge.

According to Dr.Hurd, the calculator is a general-purpose data processing machine not designed specifically for translating. Nevertheless, it has a memory capable of storing roughly 1,000,000 five-letter words. There are 600,000 entries in the latest Webster's unabridged New International Dictionary.

Dr.Hurd said that the corporation would now design a machine particularly fit for translating rather than for general computing utility. Such a device should be ready within three to five years, when the Georgetown scholars believe they can complete the "literary" end of the system.

Dr.Dostert and Dr.Garvin said they chose Russian for their first experiments because it was a difficult language and a system that could translate it could handle anything.

The machine will not accept incoherent statements, Dr.Dostert said. If they are introduced for "translation" the machine balks, and rings a bell. And it will ring a bell when it encounters a misprint. It now prints eighty letters in two seconds.

As soon as cards for Russian are completed, sets will be made for German and French. Then other Slavic, Germanic and Romance languages can be set up at will.>>

A further example of the impact of the demonstration can be found in the *Christian Science Monitor* a few days later (11 January 1954), a report which possibly reflects more accurately the exaggerated popular expectations of computers at the time.

<< **Robot translates nimbly** by Harry C.Kenny

Kachyestvo uglya opryedyelyayetsya kaloryiynostjyu, i.e. The quality of coal is determined by calory content. Thus, the Soviet language, has been translated into English by an electronic "brain" for the first time.

Yet so far as the now famous International Business Machines computer 701 is concerned, it doesn't blink an additional red light or add another whir to its purr if another language is inserted into the machine to come out in seconds in English. The Soviet language was chosen at the IBM demonstration here only because there is a relatively small number of students here of the Soviet language as it is so difficult to translate.

Significance pointed up.

The significance of the machine also is pointed up by the fact that there is a steadily growing accumulation of Soviet textual material whose true significance cannot even be estimated until its content can be converted into English.

The girl who operated 701 did not understand a word of Soviet speech and yet more than 60 Soviet sentences were given to the "brain" which translated smoothly at the rate of about 2½ lines a second. And then just to give the electronics a real workout, brief statements about politics, law, mathematics, chemistry, metallurgy, communications, and military affairs were submitted in the Soviet language by linguists of the Georgetown University Institute of Languages and Linguistics.

Flicked out nonchalantly.

The "brain" didn't even strain its superlative versatility and flicked out its interpretation with a nonchalant attitude of assumed intellectual achievement.

It is expected by IBM and Georgetown University, which collaborated on this project, that within a few years there will be a number of "brains" translating all languages with equal aplomb and dispatch.

"The potential value of this experiment for the national interest in defense or in peace is readily seen," Prof. Leon Dostert, Georgetown language scholar, said. Professor Dostert originated the practical approach to the idea of electronic translation. Along with Dr.Paul Garvin, director of the translation project, he spoke to the group of natural scientists, United States Government specialists, and the press who witnessed the demonstration at IBM headquarters here on Madison Avenue.

Enormous Potential.

"Those in charge of this experiment," the professor continued, "now consider it to be definitely established that meaning conversion through electronic language translation is feasible." Although he emphasised it is not yet possible "to insert a Russian book at one end and come out with an English book at the other", the professor forecast that "five, perhaps three, years hence, interlingual meaning conversion by electronic process in important functional areas of several languages may well be an accomplished fact."

Actually, this demonstration was rated only as a scientific sample, or, as Professor Dostert put it, "a Kitty Hawk of electronic translation." Nevertheless, the success of the project contains enormous implications for both linguistics and electronics. It is expected by IBM officials that the day will arrive when a simpler and cheaper machine will be available for less than the present $500,000 supercalculator.... >>

Two days later (13th January 1954), the *Christian Science Monitor* carried an editorial on the demonstration - one of the very few editorial statements about MT ever - which puts it into context. While the two reports of the demonstration emphasised the imminence of translation machines, the editorial gave a more sober (and, in hindsight, more realistic) assessment:

"Such an accomplishment, of course, is far from encompassing the several hundred thousand words which constitute a language. And with all the preparations for coping with syntax, one wonders if the results will not sometimes suggest the stiffness of the starch mentioned in one of the sentences as being produced by mechanical methods. Nevertheless, anything which gives promise of melting some of the difficulty which writers and speakers of different languages encounter in understanding each other -particularly as between English and Russian today - is certainly welcome."

A fascinating account of the technical problems which had to be surmounted was given by Peter Sheridan ('Research in language translation on the IBM type 701') in the *IBM Technical Newsletter* no.9, published in January 1955. Every aspect of the process sent the programmers into unknown territory: they had to decide about the coding of alphabetic characters, how the Russian letters were to be transliterated, how the Russian vocabulary was to be stored on the magnetic drum, how the 'syntactic' codes were to operate and how they were to be stored, how much information was to go on each punched card, etc. Detailed flow charts were drawn up for what today would be simple and straightforward operations, such as the identification of words and their matching against dictionary entries. The problems described by Sheridan are an illuminating illustration for younger MT researchers of the nature of the difficulties faced by the computer pioneers, who had to program with no 'assembly language' codes, let alone 'high-level' programming languages.

The linguistic aspects of the experiment were given in a contemporary account by Dostert himself ('The Georgetown-I.B.M. experiment' in Locke, W.N. and Booth, A.D. (eds.) *Machine translation of languages* (Cambridge, Mass.: M.I.T.Press, 1955), pp.124-135). Dostert had been a participant of the first MT conference two years earlier, held at MIT in June 1952 [see MTNI#2 (May 1992), 11-12]. He had come away convinced that "rather than attempt to resolve theoretically a rather vast segment of the problem, it would be more fruitful to make an actual experiment, limited in scope but significant in terms of broader implications". He identified "the primary problem [as] one of linguistic analysis, leading to the formulation in mechanical terms of the bilingual transfer operations, lexical or syntactic." The aim was a system requiring no pre-editing of the input, and producing "clear, complete statements in intelligible language at the output", although "certain stylistic revisions may...be required..., just as when the translation is done by human beings."

The corpus of 49 Russian sentences with a lexicon of just 250 words had been carefully selected. The lexical items were coded in a system of "digital diacritics" of three types. One series of 'diacritics' (Program Initiating Diacritics) indicated which of the six operational rules was to be applied. The second series (Choice Determining Diacritics) indicated what contextual information should be sought to determine selection of output. And the third (Address Diacritics) indicated the storage area of the English equivalents.

The six operational rules were (using the numbering of the researchers):

0. No problems of selection: there is one-to-one equivalence of source and target words, and the word order of the source is to be followed.

1. There is a change of order: the words are to be inverted.

2. The choice between target equivalents is determined by an indication ('diacritic') in the following word

3. The choice of target words is determined by an indication in the preceding word

4. The word in the source is omitted, and no word appears in the target sentence

5. A word is inserted in the target which has no correspondent in the source.

The limitations of the experiment were freely admitted in a later evaluation by Paul Garvin ('The Georgetown-IBM experiment of 1954: an evaluation in retrospect' *Papers in linguistics in honor of Léon Dostert* (The Hague: Mouton, 1967), pp.46-56; reprinted in his *On machine translation* (The Hague: Mouton, 1972), pp.51-64). The limitations were the consequence of restricting the algorithm to "a few severely limited rules, each containing a simple recognition routine with one or two simple commands." Nevertheless, in Garvin's view, the experiment was "realistic because the rules dealt with genuine decision problems, based on the identification of the two fundamental types of translation decisions: selection decisions and arrangement decisions." The limitations were of two principal types: the restriction of the search span to immediately adjacent items, the restriction of target words to just two possibilities, and the restriction of rearrangements to two immediately adjacent items.

The choice of target language equivalents was restricted to those which were idiomatic for the 49 sentences only. The limitation of the procedure for Russian case endings was severe: either a case suffix was not translated at all or it was translated by one "arbitrarily assigned" English preposition. Further limitations were highlighted by Michael Zarechnak, a member of the Georgetown MT group, in his essay 'The history of machine translation' in: Bozena Henisz-Dostert, R.Ross Macdonald and Michael Zarechnak *Machine translation* (The Hague: Mouton, 1979), pp.20-28. None of the Russian sentences had negative particles; all were declaratives; there were no interrogatives or compound sentences (coordinate or subordinate clauses); the verbs were all in the third person; and English articles were inserted arbitrarily to fit the particular words of the corpus.

Such limitations made it possible for the output to be impressively idiomatic, and suggested that continued experiments on the same lines would lead to systems with larger vocabularies and even better quality. The general public was impressed: MT was now seen as a feasible objective, and the translation quality was certainly acceptable. The demonstration undoubtedly encouraged US government agencies to support research on a large scale for the decade, and it stimulated the establishment of MT groups in other countries, notably in the USSR.

On the other hand, unrealistic expectations had been raised: the newspapers reported Dostert's optimism about working systems in the next three to five years, and with even better performance. They did not materialize. The disappointment led in the end to the ALPAC report of 1966, which brought the large-scale funding of MT research in the United States to a virtual end. Indeed, the ALPAC report used the public achievements of the Georgetown-IBM experiment as a touchstone when assessing the output quality of subsequent systems. In doing so, it failed to acknowledge the artificiality of this small-scale demonstrator system.

In later years, MT researchers have been rather more circumspect when demonstrating experimental systems and have been less willing to indulge in speculations for journalists. The painful lessons of the Georgetown-IBM demonstration seem to have been learned.

PEOPLE ON THE MOVE...

Winfield Scott Bennett has been promoted to Director, Linguistic Development, at Logos Corporation. Scott can be reached at Logos, 200 Valley Road, Suite 400, Mt. Arlington, NJ 07856; tel. 201 398 8710, fax. 201 398 6102.

Sergei Nirenburg will be leaving Carnegie Mellon University and joining the faculty of New Mexico State University, where he has been appointed Professor of Computer Science and Director of the Computing Research Laboratory (New Mexico State University, Computing Research Laboratory, Box 30001, Las Cruces, NM 88003; email: sergei@nmsu.edu).

Roll Lolling is retiring from his post at DG XIII of the Commission of the European Communities. His responsibilities for the organization of MT Summit V have been assumed by Sergei Perschke.

NEW PUBLICATIONS

Bibliographies in Computational Linguistics from INFOLINGUA

Conrad Sabourin has founded *Infolingua Inc.* to publish a series of extensive and fully indexed bibliographies, which he has compiled over the past years. The database for the bibliographies has been created by scanning systematically 400 periodicals over 40 years and 800 conference proceedings, covering the whole spectrum of natural language processing and related fields. Whenever a document was identified for inclusion, its references were also checked for potential inclusion. By the end of 1993, coverage was considered complete enough to justify publication. Seventeen subfields with between 2300 and 8300 references each are now being published in a collection of 25 volumes. Because some references belong to more than one subfield, approximately 15% of the references can be found in more than one volume. This does mean, however, that each volume covers in itself a complete subfield.

Conrad Sabourin is considering the publication of updates, which may well include introductory surveys, state-of-the-art reviews, and historical descriptions of the relevant fields. The continuation of this valuable project will depend on the interest of the research communities concerned, and the editor invites comments, suggestions and of course additions.

Of direct relevance to readers of MT News International is the bibliography for *Machine translation*. In two volumes, this publication lists over 8000 citations to articles and books dealing with MT, machine aids and tools for translators, translators' workstations, speech translation, etc. The main body of the work contains full bibliographic references arranged by authors and subdivided by date of publication, covering from 1950 to 1993. The index provides access to articles (via the numbers of entries in the main part) from languages analyzed or generated, from model types, system modules (e.g. dictionaries grammatical features and theories, system name, and much more. As the most complete bibliography to date in the field this provides an essential tool for all MT centres worldwide.

Details of these two volumes and the others in the series, together with information about ordering and payment are given below.

NATURAL LANGUAGE INTERFACES: Interfaces to Databases, to Expert Systems, to Robots, to Operating Systems, and to Question-
Answering Systems: BIBLIOGRAPHY, by Conrad F. SABOURIN. 1994, 2 volumes, 847p, ISBN=2-921173-08-5,2-921173-09-3 prepaid US$ 130. Total Number of references: 4100

MACHINE TRANSLATION: Aids to Translation, Speech Translation: BIBLIOGRAPHY, by Conrad F. SABOURIN and Laurent R. BOURBEAU. 1994, 2 volumes, 1168p, ISBN=2-921173-10-7, 2-921173-11-5 prepaid US$ 180. Total Number of references: 8070

COMPUTATIONAL MORPHOLOGY: Morphological Analysis and Generation,
Lemmatization: BIBLIOGRAPHY, by Conrad F. SABOURIN. 1994, 492p, ISBN=2-921173-01-8 prepaid US$ 80. Total Number of references: 2350

COMPUTATIONAL PARSING: Syntactic Analysis, Semantic Analysis, Semantic Interpretation, Parsing Algorithms, Parsing Strategies: BIBLIOGRAPHY, by Conrad F. SABOURIN. 1994, 2 volumes, 1029p, ISBN=2-921173-02-6, 2-921173-03-4 prepaid US$ 150. Total Number of references: 5180

COMPUTATIONAL LEXICOLOGY AND LEXICOGRAPHY: Dictionaries, Thesauri, Term Banks; Analysis, Transfer and Generation Dictionaries; Machine Readable Dictionaries; Lexical Semantics; Lexicon Grammars: BIBLIOGRAPHY, by Conrad F. SABOURIN. 1994, 2 volumes, 1031p, ISBN=2-921173-04-2, 2-921173-05-0 prepaid US$ 150. Total Number of references: 5910.

COMPUTATIONAL TEXT UNDERSTANDING: Natural Language Programming, Argument Analysis: BIBLIOGRAPHY, by Conrad F. SABOURIN. 1994, 657p, ISBN=2-921173-06-9 prepaid US$ 80. Total Number of references: 3830

COMPUTATIONAL TEXT GENERATION: Generation from Data or Linguistic Structure, Text Planning, Sentence Generation, Explanation Generation: BIBLIOGRAPHY, by Conrad F. SABOURIN with a survey article by Mark T. Maybury. 1994, 649p, ISBN=2-921173-07-7 prepaid US$ 80. Total Number of references: 2870

LITERARY COMPUTING: Style Analysis, Author Identification, Text Collation, Literary Criticism: BIBLIOGRAPHY, by Conrad F. SABOURIN. 1994, 581p ISBN=2-921173-12-3 prepaid US$ 80. Total Number of references: 4060

COMPUTER ASSISTED LANGUAGE TEACHING: Teaching Vocabulary, Grammar, Spelling, Writing, Composition, Listening, Speaking, Translation, Foreign Languages; Text Composition Aids, Error Detection and Correction, Readability Analysis: BIBLIOGRAPHY, by Conrad F. SABOURIN and Elca TARRAB. 1994, 2 volumes, 1066p, ISBN=2-921173-13-1,2-921173-14-X prepaid US$ 150. Total Number of references: 8010

COMPUTER MEDIATED COMMUNICATION: Computer Conferencing, Electronic Mail, Electronic Publishing, Computer Interviewing, Interactive Text Reading, Group Decision Support Systems, Idea Generation Support Systems, Human-Machine Communication, Multi-Media Communication, Hypertext, Hypermedia, Linguistic Games: BIBLIOGRAPHY, by Conrad F. SABOURIN and Rolande M. Lamarche. 1994, 2 volumes, 862p, ISBN=2-921173-15-8, 2-921173-16-6 prepaid US$ 130. Total Number of references: 5680

ELECTRONIC DOCUMENT PROCESSING: Document Editing, Formatting, Typesetting, Coding, Storing, Interchanging, Managing: BIBLIOGRAPHY, by Conrad F. SABOURIN. 1994, 551p, ISBN=2-921173-17-4 prepaid US$ 80. Total Number of references: 4260

COMPUTATIONAL CHARACTER PROCESSING: Character Coding, Input, Output, Synthesis, Ordering, Conversion; Text Compression, Encryption, Display; Hashing; Literate Programming: BIBLIOGRAPHY, by Conrad F. SABOURIN. 1994, 580p, ISBN=2-921173-18-2 prepaid US$ 80. Total Number of references: 4120

QUANTITATIVE AND STATISTICAL LINGUISTICS: Frequencies of Characters, Phonemes, Words, Grammatical Categories, Syntactic Structures; Lexical Richness, Word Collocations, Entropy, Word Length, Sentence Length: BIBLIOGRAPHY, by Conrad F. SABOURIN. 1994, 508p, ISBN=2-921173-19-0 prepaid US$ 80. Total Number of references: 3100

MATHEMATICAL AND FORMAL LINGUISTICS: Grammar Formalisms, Grammar Testing, Logics, Quantifiers: BIBLIOGRAPHY, by Conrad F. SABOURIN. 1994, 612p, ISBN=2-921173-20-4 prepaid US$ 80. Total Number of references: 3840

COMPUTATIONAL SPEECH PROCESSING: Speech Analysis, Recognition, Understanding, Compression, Transmission, Coding, Synthesis; Text to Speech Systems, Speech to Tactile Displays, Speaker Identification, Prosody Processing: BIBLIOGRAPHY, by Conrad F. SABOURIN. 1994, 2 volumes, 1187p, ISBN=2-921173-21-2, 2-921173-22-0 prepaid US$ 150. Total Number of references: 8290

COMPUTATIONAL LINGUISTICS IN INFORMATION SCIENCE: Information Retrieval (Full-Text or Conceptual), Automatic Indexing, Text Abstraction, Content Analysis, Information Extraction, Query Languages: BIBLIOGRAPHY, by Conrad F. SABOURIN. 1994, 2 volumes, 1047p, ISBN=2-921173-23-9, 2-921173-24-7 prepaid US$ 150. Total Number of references: 6390

OPTICAL CHARACTER RECOGNITION AND DOCUMENT SEGMENTATION: Character Preprocessing, Thinning, Isolation, Segmentation, Feature Extraction; Cursive and Multi-Font Recognition, Writer/Scriptor Identification: BIBLIOGRAPHY, by Conrad F. SABOURIN. 1994, 512p, ISBN=2-921173-25-5 prepaid US$ 80. Total Number of references: 3700

ORDERING INFORMATION

Payment: All orders must be prepaid in U.S. dollars, by:
- Bank draft drawn on a U.S. bank
- or: International money order

Shipping by surface mail: free
by air mail: inside North America: US$ 5.00 per vol.
outside North America: US$ 12.00 per vol.
Sales tax (Canadian residents only): add 7% GST

Payable to: INFOLINGUA Inc., P.O. Box 187 Snowdon, Montreal, Qc, H3X 3T4, Canada
Further information - Email: 73651.2144@compuserve.com

Reusing Grammatical Resources

The fourth volume of the series *Studies in Machine Translation and Natural Language Processing* published by the Commission of the European Communities is devoted to two closely related topics of considerable current interest in MT research: the reusability of grammatical resources and the efficient encoding of grammatical descriptions. The context of the research reported is stated in the introduction: "substantial computational grammatical resources exist in various NLP systems, and large scale descriptions must be quickly produced if applications are to succeed... in the CL community, there is a perceptible paradigm shift towards typed feature structure and constraint based systems and, if successful, migration allows such systems to be equipped with large bodies of linguistic descriptions drawn from existing resources." Migration refers to the transfer of linguistic resources from one computational formalism to another. The possibilities are explored with various formalisms as 'source': LFG, HPSG and the Eurotra ETS formalism, and with the ALEP_0 formalism as 'target' (ALEP was developed from the ET-6.1 formalism of the Eurotra project.) In addition, the volume contains papers on issues of efficient and 'natural' grammatical descriptions in unification-based formalisms. The volume has been edited by Stella Markantonatou and Louisa Sadler, both of the Department of Languages and Linguistics at the University of Essex (U.K.)

The previous (third) volume in the series - published late last year - is devoted to 'Preference in Eurotra' and has been edited by Paul Bennett and Patrizia Paggio. The volume represents a rich source of information about the formalism of semantic rules in Eurotra and about the theories which have influenced the researchers.

Details of the articles in both these volumes appear in the section 'Publications received' below.

Report published on Verbmobil Project

Martin Kay, Jean Mark Gawron and Peter Norvig have written a book published in the CSLI Lecture Notes series on the *Verbmobil* project [for details see 'Publications Received']. The aim of Verbmobil is to build a portable simultaneous interpretation machine. This ambitious project sponsored by the German Federal Ministry of Research and Technology demands solutions to major problems in speech recognition, the language of everyday conversation and, of course, machine translation. The book is not a report of progress on Verbmobil as such, since it was written before the beginning of the project, nor is it a feasibility study for Verbmobil since the decision to undertake the research had already been made. Rather it summarizes the research foundations; it contains the authors' assessment of "the state of the art in relevant fields of science and technology and to chart a course to the first prototype". The first third of the book consists of a review of research on machine translation (pp.11-106) - its problems, system types, the linguistic issues, translation strategies, and a survey of systems up to circa 1991. The second third is a review of research on speech recognition (pp.109-157) - its problems, the approaches, problems of synthesis and prosody, and a survey of current systems. The final third (pp. 161-211) outlines the authors' recommendations in detail. These are, in brief, that the Verbmobil project should aim to deliver two products by the year 2000: (a) a system for "face-to-face conversation between a pair of participants, each speaking a different language, on extremely limited subject matter, and in circumstances in which the conversational aims of the participants are known in advance"; and (b) a system for "the interpretation of occasional words and phrases in a conversation between a pair of participants who communicate mainly in a common language." It is essentially the second of these products which is being currently explored by the German researchers. In view of the potential importance of the Verbmobil project for the future public image of MT in general it is clear that this book must be required reading for the MT research community.

PUBLICATIONS RECEIVED

Journals

AAMT Journal: Journal of the Asia-Pacific Association for Machine Translation, *no.5 (December 1993).* p.1: Expectations of machine translation (Eiichi Ohno). -- p.2-3: MT in New Zealand: one researcher's perspective (Minako O'Hagan). -- p.4-5: The Malaysian National Institute of Translation (Ahmad Zaki Abu Bakar). -- p.6: Communication and information systems research laboratories at Research and Development Center, Toshiba Corporation (Shin-ya Amano). -- p.7-9: The limitless world of translation (Tetsuo Sagara). -- p.10-12: Interlingua in machine translation (Shin-ichiro Kamei). -- p.14-15: Introduction to IBM TranslationManager/2. -- p.16-17: Japanese-English translation software TransLand [see ?? this issue]. -- p.18-19: LogoVista E to J Macintosh version V2.0.
no.6 (March 1994). p.1: The profound world of translation (Sadao Hoshino). -- p.2-3: An introduction to the LED of CS&S [China National Software and Service Corporation] (Guan Weizhong). --p.5-9: How to utilize machine translation in a commercial translation house [Nagase & Co.] (Susumu Donomae). -- p.10-19: The JICST frequency dictionary bilingual corpora (Tatsuo Ashizaki). -- p.20-21: Fujitsu Laboratories Ltd. (Kenji Sugiyama). -- p.22-25: Future trend of machine translation technology (Hitoshi Isahara). -- p.26-27: DUET Qt version 2.1 by Sharp Ltd.

Computational Linguistics, *vol.19, no.3 (September 1993).* p.409-449: Evaluating message understanding systems: an analysis of the Third Message Understanding Conference (MUC-3) (Nancy Chinchor, Lynette Hirschman, and David D.Lewis). -- p.451-499: A computational theory of goal-directed style in syntax (Chrysanne DiMarco and Graeme Hirst). -- p.501-530: Empirical studies on the disambiguation of cue phrases (Julia Hirschberg and Diane Litman). -- p.531-538: Co-occurrence patterns among collocations: a tool for corpus-based lexical knowledge acquisition (Douglas Biber). -- p.539-564: Book reviews.
vol.19, no.4 (December 1993). p.571-590: The interface between phrasal and functional constraints (John T.Maxwell and Ronald M.Kaplan). -- p.591-636: Parsing some constrained grammar formalisms (K.Vijay-Shanker and David J.Weir). -- p.637-650: Indexical expressions in the scope of attitude verbs (Andrew

R.Haas). -- p.651-694: Planning text for advisory dialogues: capturing intentional and rhetorical information (Johanna D.Moore and Cécile L.Paris). -- p.695-704: Book reviews.
vol.20, no.1 (March 1994). p.iii-vi: Donald E.Walker: a remembrance (Barbara Grosz and Jerry R.Hobbs). -- p.1-25: Computing with features as formulae (Mark Johnson). -- p.27-54: A hierarchical stochastic model for automatic prediction of prosodic boundary location (M.Ostendorf and N.Veilleux). -- p.55-90: One-level phonology: autosegmental representations and rules as finite automata (Steven Bird and T.Mark Ellison). -- p.91-124: An alternative conception of tree-adjoining derivation (Yves Schabes and Stuart M.Shieber). -- p.125-154: Book reviews.

Language Industry Monitor *no.19 (January-February 1994).* p.1-4: The spirit of Eurotra [see this issue]. -- p.5-6: Winger: still hanging in there [see this issue].
no.20 (March-April 1994). p.1-3: Bonjour, Eurolang Optimizer. --p.9-11: LogoVista conquers Japan. -- p.11-12: Systran flourishes.

Language International, *vol.5 no.6 (December 1993).* p.11-13: Attitudes towards machine translation (Siety Meijer). -- p.13-14: Whither MT? (Colin Brace) [Report of TMI-93 conference, Kyoto].
vol.6 no.1 (February 1994). p.6-7: The pragmatic approach gaining ground [Report of Translating and the Computer conference.]

LISA Forum Newsletter, *vol.3 (February 1994):* p.1: Safety in numbers (Colin Brace). -- p.2: The LISA annual meeting. -- p.3-12: Future translation workbenches: some essential requirements (Richard Ishida). -- p.12-15: IBM's TRanslationManager: a summary of functions (Edward Lippmann). -- p.15-16: TRADOS' TAlign (Matthias Heyn). -- p.16-18: TRANSIT for Windows (Petro Antonicelli). -- p.18-19: The Joust^{TM} translation workstation (Siu-Ling Koo). -- p.21-29: The current state of MT usage. Or: how do I use thee? Let me count the ways? (Muriel Vasconcellos). -- p.29-31: Upward mobility - EUROLANG's METAL portion (Thomas Schneider). -- p.32: An integrated workplace for translators and terminologists (Hubert Lehmann). -- p.33: Logos' linguistic and system strategy. -- p.34-36: The LISA Executive Committee reports to the General Assembly.

Literary & Linguistic Computing, *vol.8 no.4, 1993.* Contents include: p.227-234: User needs for textual corpora in natural language processing (John McNaught). -- p.235-242: Developing effective resources for research on texts, using texts, and putting texts in context (Susan Hockey and Donald Walker). --p.243-258: Representativeness in corpus design (Douglas Biber). -- p.259-266: Spoken corpus design (Steve Crowdy). -- p.267-274: The need for grammatical stocktaking (Geoffrey Sampson). --p.275-282: Corpus annotation schemes (Geoffrey Leech).

Machine Translation, *vol.8, no.4, 1993:* p.209-235: Generation of text from logical formulae (John D.Phillips). -- p.239-258: Good applications for crummy machine translation (Kenneth W.Church and Eduard H.Hovy). -- p.259-279: Book reviews (Peter G.Peterson, A.P.Neal, Doug Arnold, Blaise Nkwenti-Azeh).

Translation News: Systran Natural Language Translation Software & Services, *vol.3 no.1* (Winter 1993-94) [see items in this issue.]

Tribune des Industries de la Langue, *no.13-14 (1993).* p.3-11: L'heure des grands travaux (André Abbou) [CEC Eurêka projects in language engineering]. -- p.12-16: Technologies vocales (Françoise Néel and Joseph Mariani). -- p.17-36: [Surveys of speech research and products in Europe, US and Japan]. -- p.37-41: Le développement des systèmes et des circuits d'information éléctronique (André Abbou and Laurence Martinet) [Teleworking in France]. -- p.42-44: Un fonds de commerce de 20,000 fournisseurs de services... Et après? (Dominique Perrin) [France Telecom].

Books

Machine translation, aids to translation, speech translation. Bibliography, [by] Conrad F.Sabourin with the collaboration of Laurent R.Bourbeau. 2 vols. (Infolingua 7.1-2) Montreal/Hudson: Infolingua, 1994. ISBN 2-921173-10-7. [For further details and information about others in the series, see 'New Publications'.]

Studies in Machine Translation and Natural Language Processing, vol.3: Preference in Eurotra. Edited by Paul Bennett and Patrizia Paggio. Bruxelles/Luxembourg: Commission of the European Communities, 1993. (ISSN 1017-6568). **Contents:** p.7-11: Introduction (Paul Bennett and Patrizia Paggio). -- p.13-25: The Eurotra preference mechanism (Patrizia Paggio). -- p.27-35: Some preference rules for English (Paul Bennett). -- p.37-48: Applying preference at the relational level (Patrizia Paggio). -- p.49-64: Lexical semantics and preference (Bolette Sandford Pedersen). --p.65-74: Preference and complex words (Heinz-Dieter Maas). -- p.75-90: Weak constraints and preference rules (Luca Dini and Giovanni Malnati). -- p.91-103: A comparative evaluation of the Eurotra preference mechanism.

Studies in Machine Translation and Natural Language Processing, vol.4: Grammatical formalisms - issues in migration. Edited by Stella Markantonatou and Louisa Sadler. Bruxelles/Luxembourg: Commission of the European Communities, 1994. (ISSN 1017-6568). **Contents:** p.7-9: Introduction. -- p.11-33: Reusability - general considerations (Douglas Arnold). -- p.35-59: Expressivity of lean formalisms (Steven Pulman). -- p.61-69: The ALEP_0 formalism (Josef van Genabith, Stefan Momma). -- p.71-90: Importation of an LFG grammar into the ALEP_0 formailsm (Dieter Kohl, Stefan Momma). -- p.91-115: English HPSG in ALEP_0 (josef van Genabith, Stella Markantonatou, Louisa Sadler, Marc Verhagen). -- p.117-137: The reuse of ET resources (Josef van Genabith, Paul Schmidt). -- p.139-165: The Spanish grammar (Toni Badia). --p.167-187: The German grammar (Paul Schmidt). -- p.189-199: Techniques and devices (Stella Markantonatou, Louisa Sadler). --p.201-209: Reflections on bitstrings, generalisation and negation (Josef van Genabith, Louisa Sadler). -- p.211-213: Template-driven phrase structure grammar (TPSG) (Josef van Genabith). --p.215-219 : Bibliography.

Verbmobil: a translation system for face-to-face dialog. [By] Martin Kay, Jean Mark Gawron, and Peter Norvig. (CSLI Lecture Notes no.33) Center for the Study of Language and Information, 1994. viii, 235 pp. (ISBN: 0-937073-96-2)

Conference proceedings

Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives... June 22, 1993, Ohio State University, Columbus, Ohio, USA. [ACL, 1993] 119pp. **Contents**: p.1-8: Robust bilingual word alignment for machine aided translation (Ido Dagan, Kenneth W.Church, William A.Gale). -- p.9-19: Robust text processing in automated information retrieval (Tomek Strzalkowski). -- p.20-29: Document filtering using semantic information from a machine readable dictionary (Elizabeth D.Liddy and Woojin Paik). -- p.30-39: Towards a cross-linguistic tagset (Jan Cloeren). -- p.40-47: HMM-based part-of-speech tagging for Chinese corpora (Chao-Huang Chang and Cheng-Der Chen). -- p.48-57: NPtool, a detector of English noun phrases (Atro Voutilainen). -- p. 58-64: Structural ambiguity and conceptual relations (Philip Resnick and Marti A.Hearst). -- p.65-73: Text recognition and collocations and domain codes (T.G.Rose and L.J.Evett). -- p.74-83: Extraction of V-N-collocations from text corpora: a feasibility study fro German (Elisabeth Breidt). --p.84-93: Computation of word associations based on the co-occurrence of words in large corpora (Manfred Wettler and Reinhard Rapp). -- p.94-101: Corpus-based adaptation mechanisms for Chinese homophone disambiguation (Chao-Huang Chang). --p.102-112: Example-based sense tagging of running Chinese text (Xiang Tong, Chang-ning Huang, Cheng-ming Guo). -- p.113-119: Experiences about compound dictionary on computer networks (Kyoji Umemura, Akihiro Umemura, Etsuko Suzuki)

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.