MT News International

Newsletter of the International Association for Machine Translation

ISSN 0965-5476 Issue no.9, September 1994

IN THIS ISSUE:

Systems and Products
Reports of Meetings
Proposals for Evaluation
Readers' Forum
Obituary: Paul Garvin
Publications and Databases
Publications Received
Conference Announcements
Forthcoming Events
Advertisements
Application and Registration Forms
Notices

Published in the United States, with the generous assistance of Microsoft Corporation

SYSTEMS and PRODUCTS

Globalink and MicroTac Sign Acquisition Agreement

[Press Release]
Globalink, Inc. announced on 11 August that it had finalized its agreement to acquire San Diego-based MicroTac Software. The combination of the two companies, expected to be completed

by mid-October, will make Globalink the single largest provider of foreign language translation software.

As a part of the acquisition, Globalink announced a new management structure, effective immediately, in which Harry E. Hagerty, Jr., currently Globalink chairman, will also become chief executive officer. Dominic A. Laiti, founder and former chief executive officer of Globalink, Inc., will assume new responsibilities as an executive consultant to market the company's new portable language products, and to establish trade opportunities with Globalink's Hong Kong distribution partner, Group Sense Limited. William E. Gregory, Jr. continues in his role as chief operating officer. Other executives include Michael Tacelosky, MicroTac president at the time of the acquisition, who will join the Globalink management team upon completion of the acquisition as chief technology officer, and Jorge Forgues, who will continue as chief financial officer.

Globalink's Power Translator and Power Translator *Professional* combined with the addition of MicroTac's Language Assistant™ series provides the market with a wide choice of translation products for Spanish, French, German, Italian, Russian, and Chinese on an equally wide range of operating platforms -- DOS, Windows™, Macintosh®, OS/2®, and UNIX®. The result is a complementary product line that will enable Globalink to offer a wide range of translation tools that meet the price and functionality requirements of each segment of the market -- education, consumer, home office, business and professional.

Globalink Announces New Chinese Translation Software

[Press release]

Globalink, Inc., recently announced the addition of Chinese to its line of Power Translator® software products. Power Translator is the first product that offers users support for both simplified and traditional Chinese characters as well as a choice of phonetic systems for input and editing.

According to President Dominic Laiti, Power Translator is the most flexible and easy-to-use Chinese translation software product available today. The product comes complete with all the fonts and drivers necessary to translate English text to Chinese, and to display, edit, and print Chinese characters without requiring a separate Chinese operating system. Users can enter the original English text directly, import a file or scan in a document. The system can be used interactively to translate sentence by sentence, or in batch mode processing. To edit or modify the Chinese translation, the system offers users support for both popular phonetic systems in use today, Pin Yin or Zhu Yin, as well as the traditional BG and Big Five mathematically based codes.

According to the announcement, Power Translator utilizes a set of sophisticated algorithms that analyze each word based on its context in the full sentence. This technology, combined with a 40,000-plus dictionary of words, phrases and idioms that are completely user-modifiable, is designed to enable users to achieve the highest translation quality possible. Power Translator for Chinese is also fast: the announcements states that the system can translate up to 20,000 words an hour. The product requires only 1 MB RAM and 15 MB hard disk space with DOS 3.1+ and a VGA graphics card on an IBM PC or compatible.

"We're very proud of the excellent reviews this product received during the beta tests. Independent tests by one of our distributors in Hong Kong confirmed that Power Translator is the best Chinese translation product available on the market today," Laiti stated. "It's an absolutely essential tool for companies doing business in the East. At an average retail price $100-150, it is the most productive, cost-effective way to communicate with colleagues in the East."

Globalink has over 20 products that are used for high productivity translation in corporations, small businesses, government agencies, and educational institutions. Globalink's Power Translator is also available in Spanish, French, or German to/from English, and Power Translator Professional is available in Spanish, French, German, or Russian to /from English, running on a wide range of operating systems -- DOS, Windows™, Macintosh®, OS/2®, and UNIX®. Additionally, Globalink

offers professional language services for all modern languages through the use of its software products, manual translations, and interpreters.

For more information about Globalink and its products, call 1-800-255-5660. Globalink's address is 9302 Lee Highway (Twelfth Floor), Fairfax, Virginia 22031-1208, USA; Fax: (703) 273-3866.

CompuServe MT Service Goes Live

Mary Flanagan

CompuServe launched its machine translation service on the MacCIM support forum on 23rd August 1994. The service maintains three parallel forums, one in English, one in French and one in German using machine translation. Translations are performed bidirectionally between English-French and English-German. Message thread topology is identical across the three forums. The forums are free of connect charges.

For further information about this service contact Pierce Reid at preid@csi.compuserve.com, or at 614-457-8600 in the US.

Interleaf Translator from Systran

[From *Language International* vol.6 no.4]

In mid-May, Systran Translation Systems unveiled the brand new Systran Gloabl Workstation for Interleaf. The premier of the compact MT system was demonstrated in UNIX at the ICON'94 conference, also known as the Users Conference for the Interleaf desktop publishing system, which was held on April 17-20 in san Diego, California.

The new application represents a major leaf in the downsizing of MT software, and it is the result of conversion of the robust mainframe technology which Systran is known for. The English multi-target system for Interleaf is expected to be shipped by autumn 1994 and will include English into French, German, Spanish, Italian and Portuguese. It will allow the user to simultaneously release multi-lingual versions of their technical information.

The speed of Systran's Interleaf Translator can be attributed to the system's residence entirely within native Interleaf, allowing the translation to be made without leaving the desktop publishing environment.

This so-called ‘layered application’ optimises use of the recently developed MT tools, the best known being the revision management software. Revision management software permits the translator to reuse previously translated text, eliminating repetitive editing chores and yielding considerable time savings on revisions to previously published documents.

At the ICON'94 conference in San Diego, the Systran Global Workstation was demonstrated on a Sun Workstation. Customers were able to see the Interleaf Translator's features in operation, including the revision management software. Pages of text including graphics and tables, formatted side-by-side on the screen, showed how the operator can move the cursor through linesof text, both on the input and target language page. Where sentences or `translation units' have been previously translated and checked, the cursor will automatically skip over such portions of text on the target text page.

Systran currently offers online access to its patented Systran translation system via a client's PC and modem. The company also offers complete localisation services, both for documentation and online copy.

Kielikone System Goes into Operation

[Based on article in *Language Industry Monitor*, no.21, May-June 1994]

After several years of evaluation and joint development, Nokia Telecommunications officially installed the Kielikone machine translation system last autumn. The Kielikone company, which has close ties to the translation industry and the research world, sells hand-held dictionaries, bilingual electronic dictionaries (Finnish to/from English, German, Swedish, and French) and morphology and spelling checkers for the Finnish language. In Finland, the announcement received widespread media coverage including TV appearances and feature articles in major newspapers.

The Kielikone system is designed as a "translator's workstation". The system is domain-independent, and not 'tuned' specifically for the Nokia domain, but it has been exhaustively tested on Nokia texts. In interactive mode, the Unix-based program displays both source and target texts on-screen and highlights the sentence being translated. In addition to basic editing functions, the system offers special post-editing functions such as an optional pop-up window listing translation equivalents for a given target language term, and function keys for modifying English articles (*a/an/the* - Finnish like Russian and Japanese has no articles and causes problems for MT systems), and changing capitalization. Unknown words are transferred unchanged and embedded in the text between asterisks. The system's general Finnish-English dictionary includes some 50,000 entries. Kielikone has a dictionary interface tool under development; currently, users can only add lexical entries for nouns, but it is intended to develop facilities for users to tune the system without actually touching the grammar rules.

The system is the result of close collaboration between Kielikone and Nokia. Kielikone has built the general lexicon and Nokia developed the domain-specific lexicon for 'information technology' ported from its own already extensive term bank. It is reported that Nokia considers that the linguistic issues have been largely solved and that it is now a matter of integrating the MT system into its Customer Documentation department staffed by some twenty translators and editors, and of establishing which documents are most suited for machine translation.

Kielikone has also been installed under test at Trantex, a translation company, and at Rautaruukki Oy, a steel company. Both are evaluating the system with a view to using it in production and are partly supporting development. Additional support is coming from TEKES, a research agency of the Finnish Ministry of Trade and Industry. At present, Kielikone is looking for partners to offer an MT service to companies: a "quick and cheap" translation service.

The system began as a research project in 1982, originally to study natural language database interfaces for Finnish. The morphological analyzer was an early spin-off. The group switched to MT in 1987, in response to requests from potential customers. The research project Kielikone (Finnish for "machine language") became the company Kielikone Oy; and Nokia was a pilot customer from the beginning. The system involved four to five full-time workers from 1987 to 1992, when it reached the product development phase. The parser is principally the work of Kielikone's managing director Harri Arnola (formerly Jäppinen). It is a dependency parser which builds structures by identifying relationships to verbs: subjects, objects, etc. (as in many Japanese parsers). Only one structure is produced for each sentence, with a claimed ninety percent accuracy. Kielikone does not attempt any semantic analysis beyond dependency relations; Arnola believes that syntactic processing provides translations of sufficient quality. At the moment, there are no plans for an English-Finnish version.

For further information contact: Kielikone Oy, Vattuniemenkuja 4, PL 126, Helsinki, SF-00211 Finland. Tel: +358-0-6820-211; Fax: +358-0-6820-167; Email: kkoy@kielikone.fi

PARS-3 English-Russian-English MT system

Michael Blekhman

The Lingvistica'93 Company of Kharkov (Ukraine) has launched version 3 of its system for translating scientific, technical, business and political texts from English into Russian and from Russian into English. The system, written in Borland Pascal 7.0, runs on IBM 286 or higher, requires MS-DOS 4.0 or higher, 450K RAM or more, at least 750K on harddisk, and a monitor supporting Cyrillic characters. It is Windows compatible and can operate in multiuser mode on Novell NetWare networks. The system includes a built-in text-editor (of files up to 1.5 MB in size) - screens may be split horizontally or vertically. PARS uses the same dictionary for translation in both directions, and there are facilities for users to compile and manage their own dictionaries. Russian words entered in users' dictionaries are automatically categorized for their noun declension or verb conjugation type. From version 3.1 PARS will support texts in WordPerfect, Word for Windows, Write and AmiPro. Help screens are provided in both English and Russian. The system is sold with a 20,000 word general dictionary, a 7,000 word business dictionary, a 20,000 word dictionary for computing, and a 20,000 word dictionary covering machine building, electrical engineering and metallurgy. Additional dictionaries can be supplied on economics (48,000 words), medicine (16,000 words), microelectronics (15,000 words), and patents and trademarks (5,000 words). Further information may be obtained from: Michael Blekhman, Director, Lingvistica '93 Co., 94a Prospekt Gagarina, apt.111, Kharkov, 310140 Ukraine (Tel: (0572)-27-71-35.)

Sietec's Metal (Also) Does Russian

Klaus Schubert

[Reprinted with kind permission from *Language Industry Monitor* no. 21, May-June 1994.]

Like most things in life, market forces seem to dictate the choice of language pairs developed for machine translation systems. If that is the case, we have come a full circle, for today there appears to be a demand again for Russian MT, just like in the 1940s and 50s, when MT was first conceived.

On the 25th of April, a Russian-German prototype of the Metal machine translation system was presented in Berlin by a three-member consortium. The event took place in an early capitalist palace in a part of the city appropriately called Siemensstadt. The prototype was funded by two German government agencies, the Amt für Auslandfragen (AfA), which gathers foreign information, and the Bundesministerium für Forschung und Technlogie (BMFT).

The Russian prototype was developed by a group of eight computational linguists and computer scientists from the former East German Academy of Sciences. The group, led by Gerda Klimonow, took the plunge into the market economy as the Gesellschaft für Multilinguale Systeme (GMS), a company set up in Germany for this purpose by a Munich-based software house. Working under the auspices of Sietec, the team was able to exploit Sietec's development environment and the existing Metal modules. The AfA had selected Sietec to develop the system after a thorough, two-year evaluation of existing MT systems, with particular attention paid to their extensibility. In turn, Sietec discovered the Academy team at the Coling '90 conference in Helsinki demonstrating its PC-based German-Russian verb translation system. The prototype has a good coverage of the grammatical complexity of the language pair. With some 7,000 entries, it boasted a set of lexicons considerably larger than is usual for a prototype. It was demonstrated with texts from two domains, nuclear power and aeronautics. Naturally, the prototype was fine-tuned to these texts, and the system produced ready-to-use texts of a quality unknown to ordinary Metal users, but in no way did Klimonow pretend the system was perfect. She listed no less than eleven major grammatical obstacles that were as yet only partially resolved. Some of these are specific to Russian-German transfer, such as the interpretation of the Russian aspect category, which does not exist in German, and the generation of definite and indefinite articles in German, which are likewise unknown in Russian.

The achievement shown in Berlin is due to a combination of expertise and thoroughness not often found in the MT world. Gerda Klimonow's team has a long-standing record of work both in German and Russian grammar as well as in computational linguistics. The team has maintained fruitful cooperation with well-respected sites in Russia and the Czech Republic, where – for want of hardware – they have for many years been forced to think, design, and plan before embarking upon implementation. In their strategy, they also stick to the often neglected distinction of linguistics, computational linguistics, and implementation, which in my experience is extremely expedient to all work in language technology.

The speakers in the Berlin event had much to say about future plans. Peer van Driesten, head of GMS, argued in favor of reversing the language pair, tackling other Slavic languages, and taking other steps toward remote cultures. Gregor Thurmair of Sietec mentioned language pairs like Russian-English and Russian-French along with Polish and Ukraninan and dwelt for quite some time on the possibility of using Metal modules for non-MT purposes, such as information skimming and routing or index term generation in other languages. Stephan Bodenkamp (AFA) advised Sietec to pay more attention to the integration of the system with the customers' work-processes, emphasizing the importance of the workbench concept.

Despite the noble sentiments espoused about the future, the Russian-German Metal project is at the end of its funding, and it is uncertain whether the prototype will be developed into a marketable product. By acquiring a complete team from the Academy, Sietec has acquired expertise far above the value of the three years or so now spent in this project It would be a shame not to pursue the project. Sietec may not reach the break-even point, however, unless it develops the German-Russian system too. Only the latter truly addresses the needs of German industry rather than the information-gathering priorities of authorities who ordered this prototype.

Technological know-how and developmental impetus cannot be put into the deep-freeze. Sietec should not drop the ball – or its competitive edge will be lost.

Duet Qt version 2.1 from Sharp

[From *AAMT Journal* no.6, March 1994]

Sharp, the pioneer manufacturer of MT, keeps ahead in the development of MT systems. Since six years ago, when Sharp put an MT machine with OCR on the market, the first of its kind in the industry, we have received a lot of acclaim from our users, focusing on the easy handling and the high translation accuracy. It was three years ago when the MT system DUET Qt, the most compact and lightest in the industry, was put on the market as a great space-saver. The bi-directional Japanese-English MT system, which had long been awaited by all, was developed by Sharp two years ago. Taking fully into account the riches of advice and suggestions from our customers, we are now releasing the latest version V2.1 of our English-Japanese MT system DUET Qt, aiming more than ever at a high quality of translation and maximum function enhancement.

Improvements and dictionary expansion for better quality of translation

1. A new method of syntactic analysis. The breadth-first method of syntactic analysis has been adopted. It selects the best candidate from several analysis results. Using this method, V2.1 puts out much better translations than before.

2. The framework for prioritizing interpretation. English sentence patterns, etc., are checked with a new framework and new rules. The best pattern is sought and selected preferentially.

3. Introduction of translation equivalent prioritization. With information such as "Case pattern priority" added to the basic dictionary, more flexible English equivalent selection is made possible.

4. An expanded basic dictionary. (1) Centering mainly on verbs, field specification, type information and semantic description has been enhanced. (2) A new dictionary structure has been adopted in order to be able to select translation equivalents. The basic dictionary structure has been

upgraded, supplying detailed information, leading to a much improved equivalent selection. (3) Large numbers of words and compounds, their synonyms, antonyms and derivates have been excerpted from a large volume of evaluation sentences and added to the basic dictionary. Important personal names have also been added. (The present total is 89,000 entries.)

Function enhancement

1. Simultaneous use of user dictionaries and special dictionaries. With our previous versions, only one user dictionary or technical terms dictionary could be used; now it is possible to translate using two at a time. Prioritization is possible between dictionaries of the same category.

2. The user dictionary can now hold more entries. Up to about 80,000 words can be registered.

3. Expansion of the exchange dictionary, displaying translation equivalents and terms consulted. When the pointed word is registered in another selected dictionary, user confirms the entry and replaces it on the same screen.

4. The translation equivalent inverse display function is reinforced. Even when translation becomes fragmented into phrase unit translation, it is now possible to refer to the English-Japanese word equivalents. Moreover, the translation equivalents of a word can now be examined and memorized.

5. A new method of learning. Some information not available before, as that about postpositions, is now also automatically included in the outcome of learning.

<example> Original sentence: *I will meet him*. (Note that *meet*, originally registered as *au* in the basic dictionary, is being relearned as *mukaeru*.)

Translation: with previous version: *Watakushi wa kare ni mukaeru*; with version V2.1: *Watakushi wa kare o mukaeru*.

6. A range-specific translation function. If the range of a text section is specified, the analysis will be confined to that range. Other parts of the original text retain their correct translation, allowing the desired translation to be found in a very short time.

7. New items in mode-setting. To translate a sentence with no subject, the translation can be output either as an imperative sentence or as indicative, based on the choice of mode, as well as whether each separate sentence is to be translated on its own. Four mode-setting items in total are added and available with V2.1.

8. The field selection function. The field is selected according to the sentence's subject matter to be translated. Up to three fields can be chosen from amongst seven available fields: chemical engineering, mechanical engineering, economics, current events, information processing, electronics, and medicine.

9. Printer environment selection. The layout of the original text and translation text, as well as selecting of single or double line is available.

10. Easier mouse operation. After translating one sentence, by double-clicking the mouse one can easily call for the available translations for any word to be displayed on the screen.

Enquiries to: Sales Promotion Section, Sharp System Products Co., Ltd. +81-43-299-8302

English/Japanese systems from Nova and Toppan

[From *Language International* vol.6 no.4]

According to newspaper reports the Japanese software developer Nova Corporation is planning to release an improved version of its automatic text translation system for the Windows operating environment. The system is for translation English to Japanese and Japanese to English and retails at $1,850 for each language direction, $3,200 for the pair. [See also MTNI#8: 5-6.]

There are also reports that another Japanese company, the Toppan Printing Corporation, has developed a Japanese to English and English to Japanese machine translation system with a built-in learning feature. The system stores previously used common translations and automatically refers to

them when the phrase recurs in the original language. The system, which is expected to go on sale next year, is reported to have cost $2 million to develop.

The JICST Frequency Dictionary Bilingual Corpora

Tatsuo Ashizaki

[Extract from article in *AAMT Journal* no.6, March 1994. The original paper was presented at the Japan-US International Workshop on Computer Aided Translation, Washington, D.C., November 22-24 1993.]

Development of JICST's MT system began with the Mu project, in which JICST participated in the creation of a MT dictionary. From 1982 to 1986, the Mu project researched and developed Japanese-English/English-Japanese MT systems. With sponsorship from the Japanese government in the form of a grant for science and technology promotion, three other organizations participated in the project with JICST, namely Kyoto University, Electronic Technology General Research Center, and RIPS.

JICST has applied the results from the Mu project in its development of an operational Japanese-English MT system, undertaken between 1986 and 1991. This system was designed to produce English translations of scientific and technological abstracts published in Japanese, and we have been actively using it to compile an English DB for extraction of citations since the summer of 1990.

The JICST MT system (J/E)

One of JICST's major services involves the accumulation of abstracts in Japanese from scientific and technological papers published both within and outside of Japan. Our English DB for science and technology papers is compiled from these abstracts. More particularly, the Japanese data are translated into English by means of MT. Prior to the actual MT process, we separate those papers fro which authors have supplied titles in English and extract the desired information directly. Others, for which titles and abstracts are given only in Japanese, are placed into two groups. The first group is checked by human editors prior to processing with a pre-editor program, while manual inspection is omitted for the other group. After pre-editing, both groups are submitted to the MT system. All MT results are post-edited by hand, and added to the English DB.

Table 1: MT-using rate in JICST English DB.

	1990	1991	1992
Grand total	234258	232615	242855
Titles	147123	156511	147065
	(63%)	(67%)	(60%)
Direct	82649	70846	73448
	(35%)	(30%)	(30%)
Human	48982		
	(21%)		
MT	15492	85665	73617
	(7%)	(37%)	(30%)
Title+Abstract	87035	76104	95790
	(37%)	(33%)	(40%)
Direct	77607	59678	95790
	(33%)	(26%)	(34%)
Human	506		
	(0%)		
MT	8922	16426	13331

	(4%)	(7%)	(6%)
MT Total	24414	102091	86948
	(10%)	(44%)	(36%)

Table 1 shows the number of items added to the English DB since 1990, when DB construction was begun using MT. For the figures under the category "Title", items are not abstracted and only title and bibliographical data are supplied in English. The "Title+Abstract" category reports items for which both title and abstract are translated. Subdivision "Direct" indicates English data was supplied by the author. "Human" refers to data produced by human translation, and the figures for "MT" show the production totals by the MT system.

As shown, MT was used for 15,000 of the 147,000 titles added in 1990 without abstracts (7% of the total items), and 8,900 of 87,000 titles with abstracts (4% of the total). Of the 234,000 total items added to the English DB in its initial year, then, roughly 10% were produced by MT.

At present, English abstracts cannot be prepared for all scientific and technological papers published in Japan. Priority is given to government reports and documents that are not readily accessible to the general public. About 5,000 such titles are included in the DB. Papers on mechatronics, an area where Japan has made significant contributions, are also abstracted. A significant feature of public resource materials is their frequent reference to farm products, geographic places in Japan and other items with highly specific or uncommon names. Most fields of science and technology exhibit a similar tendency to adopt elaborate terminology. The ongoing incorporation of such terminology in our MT system has greatly enhanced the rate of MT production for JICST's English DB, as the figures for 1991 show. In the "Title" category, ratio of MT processing increased to 37% for 7% in 1990. "Title+Abstract" was 7% in 1990, compared with 4% for the previous year. With respect to the total items added to the English DB, 44% were successfully processed by MT in 1991, followed by 36% in 1992.

Table 2 shows the aggregate accumulation of titles in the English DB from its inception in 1984, prior to the use of MT, through 1992. The average is 200,000 titles annually, and the DB currently holds some two million total items.

Table 2: Yearly entry and accumulation of JICST DB

Year	Entries added	Accumulation
1984	18454	18454
1985	173862	192316
1986	200101	392417
1987	235867	628284
1988	260262	888546
1989	244068	1132614
1990	234158	1366772
1991	232615	1599387
1992	242855	1842242

Post-editing of MT output

Because many features of English cannot be fully accommodated by the MT system, manual post-editing is necessary. Such features include, e.g., number (singular/plural) inflection, preposition selection, and word order change resulting from sentence structure modification. Human post-editors typically find it most difficult to correct sentence modifications and parallel phrases, as there are still no universal post-editing rules for determining the proper tense. For example, the Japanese pattern "... ni tsuite nobeta" may be perceived as past or present tense in a sentence out of context. Because no set rules exist for correspondence between Japanese and English sentences, different people editing the same MT output based on personal judgment will often suggest different changes. Thus

the post-edited sentence is non-deterministic, and as such the quality of human post-editing is often little better than the raw MT result itself.

Fitness of MT input

Due to their inherent complexity, long sentences (i.e. those having 150 characters or more) are designated as unsuitable input to JICST's MT system. Sentences with complicated modification or lengthy hiragana strings of 10 or more characters are also problematic. Such sentences deter the compilation processes and cause the system to output Japanese sentences. It is considerably bothersome for the post-editor to proof-read texts with Japanese sentences that result from such MT failures. In essence, post-editing of MT output in Japanese would require time-consuming manual translation into English. Accordingly, at the morpheme generation stage, the MT-failed, or J/E MT output in Japanese, are forwarded again to the morpheme analysis process, but the sentence structure is not executed.

In general, the second morpheme analysis undertaken to assist the post-editors comprises a dictionary look-up for each Japanese word in the MT-failed sentence. English equivalents are listed word by word in the order of the original Japanese text, and the post-editor generates an English sentence from this output. For nouns and adjectives, the system gives the first English equivalent in the translation dictionary, while verbs and particles are left but in Japanese.

It may be noted that the average person would not understand the contents of scientific and technological abstracts even in his or her native language, given the specialized terminology involved. An expert in the associated field, however, will readily comprehend the essential terms, including nouns, adjectives and verbs, as translated by the MT system. It may thus be posited that as long as the raw MT output has enough information to convey an understanding of the meaning, post-editing is not critical, but serves mainly to smooth out the English expressions. In this sense, the current MT system can be said to be fully automated. While a single English equivalent to each Japanese expression is designated, the system lists a synonym and its corresponding English equivalent as well. This feature benefits not only the post-editors to gain a better understanding of the original text, but anyone else accessing the raw MT output as well.

Word entries in the Translation Dictionary

Table 3 gives a breakdown of the number of entries for various parts of speech in the translation dictionary. In Table 3, "NOUN (S&T)" refers to the JICST dictionary of science and technology terms, and "NOUN (MEDICAL)" to the MEID dictionary converted into JICST dictionary format. Figures in parentheses under "PROPER NOUN" refer to place names, organization names, and the like.

Table 3: Number of words in dictionary (as of July 1993)

	Japanese	J-E	English
P.O.S.			
NOUN (S&T)	352016	352160	294250
NOUN (MEDICAL)	183092	183092	119306
PROPER NOUN	(24896)	(24896)	(806)
PRONOUN	30	30	47
VERB	15256	14398	4971
ADJECTIVE	6795	6795	7088
ADVERB	451	451	2123
DETERMINER	155	155	48
CONJUNCTION	87	87	34
POSTPOSITION	102	87	
AUXILIARY VERB	115	115	9
AFFIX	279	252	
PREPOSITION			158
ARTICLE			3
CARDINAL NUMBER			30

ORDINAL NUMBER			30
UNIT			5
TOTAL	558348	557594	428097

Japanese verbs in the dictionary number about 15,000, but only some 5,000 English verbs are included. This is because the verb entries are drawn from, and thus correspond to, practical sample texts. Japanese technical writing makes frequent use of modified verbs. Upon translation, sentences incorporating such verbs are cast in patterns that better reflect English sentence structure. Consequently, when compared with the English equivalents, the number of Japanese verbs is significantly greater. An example of this verb correspondence is seen in the sentence pattern analysis of the Japanese expression *'A' o deta-shori suru*, which might be literally interpreted as *Data process 'A'*. A more natural English equivalent, however, is obtained by substituting a verb and a noun for the verb in the initial expression, viz. *'A' no deta wo shorisuru*, or *Process A's data*.

[The remainder of the article discusses other problems with Japanese compound nouns, adjectives, and proper names; and describes in detail the compilation of various dictionaries and tables: Correspondence dictionary, Translation dictionary, Part of Speech Table, Dictionary Entry Format, Japanese Deep Case, and Frequency Dictionary.]

REPORTS of MEETINGS

MT Showcase at AAAI-94

Joseph Pentheroudakis

The twelfth national conference of the American Association for Artificial Intelligence (AAAI), held in Seattle, Washington, from 2-5 August, included for the first time a showcase exhibition of MT systems. The showcase was organized by Jaime Carbonell (CMU), Bonnie Dorr (Univ. of Maryland) and Eduard Hovy (ISI). The two-day event was attended by many AAAI participants; visitors could stop in at the exhibitors' booths, or see one of the formal system presentations.

In all, ten systems were represented in the showcase: Kant, Logos, LogoVista E to J, Metal, Eurolang Optimizer, Pangloss, Princitran, Spanam/Engspan, Swetra, and Systran.

The KANT system, developed at Carnegie Mellon University's Center for Machine Translation, was presented by Alex Franz and Jaime Carbonell. KANT, an interlingua-based system, is designed for applications where technical information is created at a central location, and then translated for world-wide dissemination. A controlled source language helps eliminate vagueness and ambiguity in the source text, thus ensuring high accuracy in the output. The Center for Machine Translation is currently building a large-scale KANT application for Caterpillar, Inc. [see MTNI#4:12]; when completed, the system will translate texts covering the entire Caterpillar product line from English into eleven different languages.

Brigitte Orliac of Logos Corporation exhibited the Logos Translation System, with particular emphasis on the Logos Semantic Table. This component is described as a collection of rules that enable the Translation System to establish the meaningful relationships among the components of a given sentence. These relations are encoded by the users with the help of SEMANTHA, the system's semantic encoding tool. A demonstration of SEMANTHA using rules selected from a variety of customer texts was also given.

LogoVista E to J's English to Japanese translation was demonstrated by Alan Pollack of Language Engineering Corp. The system runs under Windows^{R} and on the Macintosh^{TM}. According to Alan Pollack, the system uses syntactic transfer with additional semantic processing to select the most appropriate translation in a given context. LogoVista E to J can be used to translate a wide variety of documents, including product manuals, technical articles, and business correspondence.

Lutz Graunitz of Sietec Open Systems showed two systems, METAL and EUROLANG OPTIMIZER. The METAL translation system, first developed at the University of Texas at Austin in the 1960's, and subsequently supported by Siemens (now Siemens Nixdorf Information Systems), is particularly well suited to high-volume draft translation of technical documentation. The system allows the addition of special technical terms to the existing dictionary, thus ensuring translator-independent consistency of translation over long periods of time and over a large variety of documents. The system also works well with several DTP files (Interleaf, Framemaker, WinWord, WordPerfect, RTF, etc.), resulting in an output document which preserves the format, graphics and tables of the original. METAL runs on several types of SPARCstations.

Additionally, Lutz Graunitz presented the EUROLANG OPTIMIZER, developed by Sonovision ITEP Technologies (SITE). The Optimizer [see also MTNI#8: 8], is a translator's workbench which includes Translation Memory and an integrated Termbank; it runs under Windows NT/MS-SQL (which functions as the server system) and WinWord 6.0, which is the client application. The system also runs under UNIX.

PANGLOSS, a joint project of CMU's Center for Machine Translation, NMSU's Computing Research Laboratory and the Information Sciences Institute at the University of Southern California, was also present at the exhibit, demonstrated by CMU's Robert Frederking. Largely in response to ARPA's MT evaluation methodology, which emphasizes fully-automatic machine translation of unrestricted newswire text, the PANGLOSS project has developed a multi-engine approach. In addition to the knowledge-based MT, the system employs a lexical transfer system as well as an example-based MT engine. Spanish to English translation was demonstrated during the exhibit; Japanese to English will eventually be added.

An experimental system for Korean to English interlingual MT, PRINCITRAN, was exhibited by Bonnie Dorr (University of Maryland) and Jye-hoon Lee (University of Manitoba). The grammatical formalisms and network representations used by the system's principle-based parser were demonstrated, as was the lexical-semantic formalism used in analyzing source language sentences; the latter will eventually serve as the interlingual input to a generator. The system is implemented in C++ and Lisp and runs on a Sun.

Marjorie Leon and Julie Aymerich of the Pan American Health Organization (PAHO) demonstrated SPANAM and ENGSPAN, translating from Spanish to (American) English and from English to Spanish, respectively. Both systems are fully automatic production systems, which together have been used to translate over 20 million words since 1980. They use augmented transition network parsers, stored in data files which can be modified at runtime. Each of the dictionaries contains approximately 65,000 lexical items, phrases, analysis rules, and lexical transfer rules. The systems are designed so that the user can fine-tune the output of the translation program by adding different types of lexical entries and context-sensitive rules to the dictionaries. The systems run on PCs. According to Marjorie Leon, the systems are not available commercially, but are being licensed from PAHO to a few public and non-profit institutions.

Developed at the Department of Linguistics at the University of Lund in Sweden, SWETRA is a research system mainly aimed at exploring solutions to theoretical linguistic problems. These problems include the relations between parsing strategies and language typology; aspect choice and article selection in translation between Slavic and Germanic languages; translation of compounds and collocations, and domain-specific dictionaries and domain-specific strategies for MT. Solutions to these issues were demonstrated by Barbara Gawronska in the context of several translation programs, implemented in LPA MacProlog and Flex on a Macintosh.

Finally, Chris Fitch of SYSTRAN presented the SYSTRAN Translation System, which last year celebrated its 25th anniversary. The company's fully-automatic translation systems comprise eleven language pairs currently used commercially (English into French, German, Spanish, Italian, Arabic, Portuguese, and Dutch; and French, German, Spanish, and Russian into English), while seventeen other language pair combinations are currently under development. SYSTRAN has evolved from a direct, mainframe-bound system translating punched-card input to a modern

transfer-type, PC-based product, handling formatted text from a variety of word processors. Future research includes integration of OCR, speech synthesis and speech understanding.
[Conference notes prepared by participating exhibitors were used in the preparation of this report.]

Arabic MT Tops AACNA Agenda

Paul Roochnik

Arabic machine translation topped the agenda at a meeting of the Association for Arabic Computing in North America (AACNA) on 24 June at Georgetown in Washington, DC. Mohammad Shihadah of AppTek, Inc., told an audience of some 25 linguists and computer scientists, that the Arabic language presents unique challenges to the implementation of an English-to-Arabic MT system. Shihadah, whose company has worked on Arabic MT research and development since 1991, characterized the design of his system as one based on the principles of Lexical-Functional Grammar (LFG):

- In the analysis stage, the system reads the English sentence one word at a time, assigning to each a set of lexical features. The system utilizes "LingWare" components (developed by Executive Communications Systems), which consist of English source language analysis, English lexicon, and linguistic formalism tools. Parsing results in two structures: a constituent structure (CS) and a functional structure (FS).

- In the transfer stage, rules modify the FS to prepare it for Arabization. Salient grammatical features are retrieved such as tense, mood, case, verb-subject agreement, etc.

- In the generation stage, the lexical entry is sought in the target dictionary and the grammatical features of the FS dictate its morphology. Arabic syntactic rules, meanwhile, dictate the word order. The system generates more than one parse tree but selects the best one, based on semantic weighting, which has developed after much experimentation on the part of AppTek linguists.

Running on a 486-33 Unix (SCO Unix: Santa Cruz Operations) platform with 32 megabytes of RAM, the AppTek system translates an average of two words per second. According to Shihadah, the company has already begun to develop Arabic-to-English MT.

AACNA began its Arabic computing forum series in March 1994 with a talk by Paul Roochnik, then of Language Analysis Systems, Inc., on the automatic processing of Arabic names. Future talks will consider such topics as Arabic optical character recognition, Arabic e-mail systems, and Arabic code standardization. Further information is available from: Jackie Murgida (jmurg@delphi.com), Musa Nasir (mnasir@guvax.georgetown.edu), Paul Roochnik (5290958@mcimail.com).

MT Workshop in Limerick

John Hutchins

On the 26th and 27th May, the University of Limerick in Ireland hosted an International Workshop on Machine Translation "Tools, resources and techniques for practical MT". The workshop was sponsored by Software and Systems Engineering Ltd. (a Siemens company) and Apple Computer, and was organised by Richard Sutcliffe (University of Limerick) and Reinhard Schaeler (University College Dublin). The first day began with a general view of the development of large-scale commercial systems given by W.Scott Bennett, based on his long experience at the University of Texas with the Metal system and his more recent experience as Director of Linguistic Development at the Logos Corporation. He stressed the practical considerations which influenced, sometimes profoundly, the design of MT systems. Bennett was followed by a critique from Reinhard Schaeler of the exaggerated claims made for all types of MT and translation tool, from fully automatic systems to translators' workbenches. He was particularly critical of vendors who fail to market honestly and do not guarantee their products. Other speakers on the first day described the use of

XL8 in translating software (Gabriele Milch-Skinner), the use of the Trados workstation (Catherine Gavin) and the problems arising in the use of the Metal system, at a Siemens subsidiary (Orla Connolly). The second day began with Yorick Wilks (University of Sheffield) on various current issues in MT: international collaboration, the debate about statistical methods, the question of interlinguas. He was followed by Matthias Heyn (Trados) who described methods for text alignment, translation memory and terminology extraction, under development for the Trados workstation, and then by Louise Guthrie (New Mexico State University), who spoke on research for automatic lexicon construction. Harold Somers presented his research at UMIST on multilingual generation, explaining that he and his colleagues were prepared to test the example-based approach to its limits. Richard Sutcliffe described the exploratory work at Limerick on using word distribution patterns in the translation context. Andrew Way (Dublin City University) discussed problems of developing test suites for MT evaluation, and the final talks were devoted to the problems of Irish lexicography and term banks (Donncha O Croínín and Gearóid Ó Néill). A distinctive feature of the workshop was the attention paid to problems of software localisation, which has become a growth industry in the Irish Republic in recent years.

Language Engineering Convention
Paris 6-7 July 1994

Organised by ELSNET and EC2 (Nanterre, France) and sponsored by the European Commission, the intention of the Convention was to present the language engineering profession to the outside world and to the practitioners themselves. An additional aim was to publicise the achievements of language engineering programmes and projects and to stimulate industrial take-up of the results. The Convention was also a forum for meetings between language engineers working in different environments, e.g. research institutes and industry. The lectures of the Convention were supplemented by an exhibition with demonstrations of current projects. Included were exhibits for the ALEP project, the DELIS prototype workbench for lexicographers, the Eurolang Optimizer and the IBM TranslationManager, and the SECC simplified English grammar and style checker.

From the published abstracts of the Convention [see 'Publications Received'] it is seen that the presentations most immediately relevant to MT included: 'ANTHEM: when machine translation meets automatic encoding' (Werner Ceusters); 'Verbmobil: speech translation on demand' (Wolfgang Wahlster); 'Eurolang: a new perspective for translation tools' (Daniel Bachut); 'Machine-assisted translation and documentation: towards the next millenium' (Khurshid Ahmad); 'TRANSLEARN: interactive corpus-based translation drafting tool' (Stelios Piperidis and George Carayannis); 'SECC: simplified English checking and correcting in an MT environment' (Gert Adriaens); and 'Evaluating translation' (Sylvie Regnier). Other talks covered software platforms and tools, localisation, controlled languages, text generation, multilingual corpora and lexica, document indexing and retrieval, system evaluation, information categorisation and extraction, technology assessment, and speech technology. In addition there were three round tables devoted to market prospects for language technology, technology trends in language engineering, and the information highway and language engineering.

The full proceedings of the Convention will be published in October, available from: Leeann Jackson-Eve, ELSNET, University of Edinburgh Centre for Cognitive Science, 2 Buccleuch Place, Edinburgh EH8 9W, Scotland (Tel: +44 31 650 4594; Fax: +44 31 650 4587; Email: elsnet@cogsci.ed.ac.uk.)

International Workshop on Sharable Natural Language Resources
Nara, 10-11 August 1994

The SNLR workshop brought together many of the leading figures in the field. In total there were one hundred participants of the workshop held by the Nara Institute of Science and Technology. The

papers given were: an introduction to the Consortium for Lexical Research (Louise Guthrie and Katherine Mitchell), the KAIST tree bank (Key-Sun Choi et al.), standards for sharing NL resources (Nicoletta Calzolari and Antonio Zampolli), improvements of the Japanese morphological analyzer JUMAN (Sadao Kurohashi et al.), the LangLAB morphological parser (Tomoyosi Akiba et al.), the CMU MT toolkit (Masaru Tomita), the KN parser (Sadao Kurohashi and Mokoto Nagao), NAIST natural language tools (Yuji Matsumoto et al.), the TFS formalism (Martin Emele), discourse tagging (Chinatsu Aone and Scott W.Bennett), Genesys (Tadashi Kumano et al.), multilingual resources in CRATER (Geoffrey Leech et al.), MULTEXT (Nancy Ide and Jean Veronis), the ECI corpus (Susan Armstrong-Warwick et al.), common syntactic dictionary of English (Catherine Macleod et al.), Japanese lexicons for computers (Minako Hasinmoto et al.), and a syntactic lexicon for English (Dania Egedi and Patrick Martin).

The workshop included poster sessions lasting up to two and a half hours during which there was the opportunity of giving several demonstrations of systems. Most of the speakers provided on-line demonstrations of the systems they had described: e.g. JUMAN, KN parser, LangLAB, CRATER, TFS, Genesys. Others giving demonstrations included Bill Ogden (X-Concord), Takahiko Kumamoto (Dialogue database), Hitoshi Isahara (ETL parser), and Ted Dunning (TkStat). Details of the published proceedings are given in `Publications Received'. They will also be available shortly through World Wide Web.

PROJECTS

China National Software and Service Corporation

Guan Weizhong

[From AAMT Journal no.6, March 1994; translated by Shravan Vasishth]

The Linguistic Engineering Department (LED) of the China National Software and Service Corporation (CS&S) was established in 1986. LED is a research and development department of CS&S. Its main task is the R&D of marketed NLP software. LED has four sub-divisions: English-Chinese machine translation, Chinese-Foreign language MT, electronic dictionaries and the Multilingual MT Project (MMT), an international joint program which was initiated by the Ministry of International Trade and Industry of Japan and organized by the Center of International Cooperation for Computerization (CICC). The department has more than 30 members involved in the above R&D programmes. Most of them are research fellows trained in linguistics, MT and software technology.

The LED has been playing an important role in the national MT development programme and in the development of NLP products in China. During the seventh Five-year plan period 1986-1990, the Chinese government invested several millions for funding the national MT programme. LED is very proud of being the organizer and coordinator for this programme. Besides this, LED also joined the international joint programme MMT as the Chinese representative, and was responsible for organizing and coordinating activities in China for the programme.

LED's research efforts have shown good results. For instance, in September 1988, LED released its first commercial English-Chinese MT system "Transtar" in China. Its intelligibility is over 60% with a processing speed of 3,000 words per hour. In July 1990, a large scale documentation database was developed for which indexing technology was provided by LED. In March 1991, a dictionary for NLP containing over 1,200 Chinese verbs was compiled. At the theoretical level, we have reviewed the abilities and deficiencies of the parser and the rules of Transtar. In addition to that, we have developed a Chinese parser and a rule description system. Several papers have been published in this regard.

Details of developments at LED are as follows:

A. English-Chinese MT system "Transtar"

English is an analytic, inflecting language. The quality of a system dealing with English depends on the depth of analysis and on the definition of logico-semantic relations between words. The parser's goal is to produce an information set reflecting the relations obtaining between words or phrases within a given sentence. While transferring, the root of the tree is taken as the kernel and the attached constituents are arranged on both sides according to a predetermined order. The LED work group has optimized the parser and the rule set in lines with the above principles. As a result, the intelligibility of the new "Transtar-92" has been improved to more than 70%, and the reading speed is now 30,000 words per hour on PC-486 machines. At present, Transtar has several hundred domestic users. It has also been sold to the US, Singapore and Hong Kong.

B. The joint MMT project

We have participated in almost all the programmes of the MMT project. LED has been the organizer on the China side, and we have undertaken the main work of Chinese text analysis and text generation. LED cooperates with Qunghua University, Northeastern University, Nanjing University and other institutions in the development of MMT. The researchers and engineers of LED study and work on the following areas: analysis, sentence generation, dictionary support, translation support, technical dictionary, and input and output from keyboard and OCR. In 1991, we sponsored the International Symposium on MMT in Beijing on behalf of the CICC. We have maintained contact with the Japan side all the time, and the atmosphere of cooperation is very satisfying. We will be making further contributions toward the completion of this project.

C. Chinese-Foreign language MT: "SinoTrans"

Generally speaking, Chinese-Foreign language MT is no different from Foreign-Chinese MT. However, the syntactic analysis of Chinese is much more difficult than that of English; using a PC-486, with SinoTrans the intelligibility of the target language is about 70% and the processing speed about 10,000 words, even after pre-editing the input text.

At present, SinoTrans has several domestic users since its announcement in September 1993.

D. Dictionaries

As is well known, an MT system works on the basis of information provided by dictionaries. The quality of output of an MT system largely depends on the integrity of its dictionaries. In this case, LED has developed more than ten English-Chinese bilingual dictionaries. They cover computer science, economics, chemical engineering and trade with the total volume amounting to 600,000 entries. As for Chinese-English, we have produced a general area dictionary of 40,000 entries. LED has also loaded the syntactic and semantic information for the Chinese-Japanese bilingual specialist dictionary. Concurrently, we are also developing electronic dictionaries for human users, and the first product is a comprehensive English-Chinese dictionary. It will be available on the market soon.

MT target in China

MT research began in China in the 1950s with Chinese and Russian as the language pairs. Work was interrupted for decades thereafter due to various circumstances, and was resumed in the late 1970s. One might say that the golden age of MT in China began from 1987. At present, MT classes are offered at several institutions. Some organizations are also carrying out research on MT. This trend is expected to continue. Mt systems other than English-Chinese (such as Japanese-Chinese, Russian-Chinese) are expected to be released in the near future.

Marketing strategy has become more and more important. Some private companies are involved in MT R&D with their own investment. Transtar, SinoTrans, Russian-Chinese System, and Chinese-Japanese System are commercially available on the market.

In the area of theoretical research, LED has caught up with the MT world. In particular, LED has acquired expertise in Chinese analysis and sentence generation. China has established a 10,000 million character corpora 50% of which will be tagged by the end of 1995. At the same time, a large-scale dictionary and an example-based tagging system are being developed.

The future of MT in China

We are keenly aware of the importance of mutual translation between foreign languages and Chinese. We are hopeful that the efforts we put into research today will bear fruit. In this connection, LED of CS&S has the following guidelines on MT R&D for the future:

1. We are actively taking part in the MMT project. We will achieve all proposed goals on the China side. We intend to open a translation service centre for the public using the improved MMT system. Commercial MT systems developed by LED will be used as well.

2. We know that MT systems are intellectual products. Their development depends on how much we know about the intrinsic nature of the concerned language. We will pay due attention to the formal grammar of Chinese in this respect. Chinese rule sets for parsing and generating will be considered as well.

3. Domestic and foreign cooperation for the compilation of new language pairs is sought; support for R&D in Chinese-German and Chinese-French systems is anticipated.

Conclusion

A quarter of the world population uses Chinese as its mother tongue. The glorious Chinese culture has been nurtured with this language. People throughout the world want to know more about China, and the Chinese people want to know more about the world. Nevertheless, parsing and generation techniques for Chinese are not achieved easily. Foreign language generation from Chinese is another serious problem. We have these problems to overcome. Therefore, we will actively take part to establish OSI with CICC's sponsorship in China. We will also continue to do our best on our side. We are very interested in mutual technical exchange in the MT circle. We believe that Chinese-Foreign language systems have a bright future.

The author Guan Weizhong is Senior Engineer/Vice-Chief Engineer and General Manager, LED, CS&S, Beijing. Fax:+86-1-8312543

The LOLITA project at Durham University

[Extracts from ELSNET listserver]

LOLITA (Large-scale, Object-based, Linguistic Interactor, Translator and Analyser), one of the most advanced natural language processing systems anywhere. Here are a few facts about LOLITA:

- based on a conceptual graph of more than 70k nodes, compatible with WordNet;
- able to perform morphological, grammatical, semantical, pragmatical and discourse analysis;
- under development for more than 8 years, at present a team of more than 20 people works on it;
- completely written in Haskell, a pure lazy functional language, with high order functions, polymorphic types and type classes (more than 35k lines of code, corresponding to about 350k lines in an imperative language);
- prototype applications include analysis of real text, NL generation, query, dialogue, template extraction, translation and language tutoring;
- processes English and Chinese; Italian, Spanish and French under development;
- very fast execution times (a parallel version under development);
- applications with Siemens Plessey, Rolls-Royce, Software AG and other major companies under development;
- chosen by the Royal Society for its prestigious 1993 Soiree Exhibition;

- registered for the 1995 MUC-6 competition (sponsored by ARPA, the Advanced Research Projects Agency of the USA); it is also going to be entered for the forthcoming SPREC and TREC competitions;

- we are among the founders of the new Journal of Natural Language Engineering, to be published by Cambridge University Press.

Areas of research are (among others): style analysis; learning of grammar & semantics rules; summarisation; metaphor and non-literals; discourse planning; semantic reasoning; large scale reasoning; rhetorics; humour; additional languages; user modelling; concept learning; emotion modelling; deep aspects of semantics, pragmatics and dialogue; concept representation; meaning correspondence between languages; integration of speech and NL; foundations of plausible reasoning etc.

For further details, write to: Dr. Roberto Garigliano, Laboratory for Natural Language Engineering, Department of Computer Science, University of Durham, South Road, Durham DH1 3LE, UK. (Email: Roberto.Garigliano@durham.ac.uk; Tel: +44 91 374 2639; Fax: +44 91 374 2560)

EVALUATION

Proposal for a Differentiated Text-Related Machine Translation Evaluation Methodology

Karin Spalink

Introduction

The evaluation of an MT system encompasses various aspects, e.g. text (input and output), hard/software compatibility, cost, user friendliness, system performance, system support, and availability of language pair combinations - not necessarily in this order. This proposal is limited to the text-related aspects of MT evaluation.

MT evaluation also has different meanings for the three parties involved, namely developers, vendors, and users. The focus of evaluation for each of these groups is very different. A vendor, for example, is probably more concerned with user friendliness than a developer, and a user might be more concerned with a specific language pair than either a developer or a vendor. This proposal approaches evaluation from a user perspective.

A text-related evaluation of machine translation from a user perspective has to consider three components: input text, MT system, and output text. Each of these components has to be evaluated in itself and in its relationship to the other components (Figure 1).

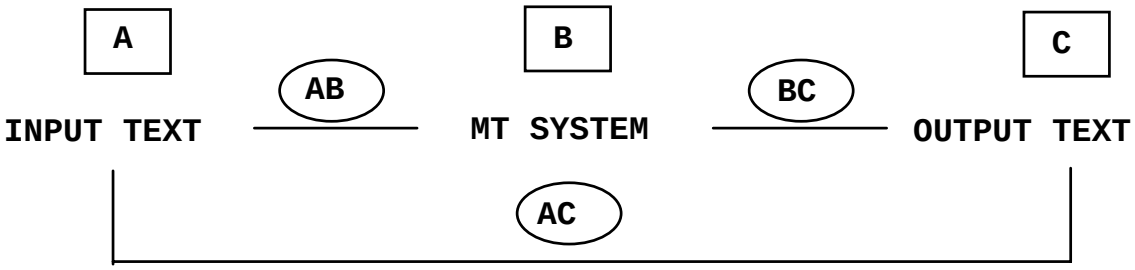

Figure 1: Components of Text-Related MT Evaluation

This differentiated evaluation procedure is especially critical in comparative evaluations, i.e. evaluation of different MT systems and evaluation of different input materials. Repetitive evaluations, i.e. evaluations of the same system at different stages, do not always require the evaluation of all components and relationships at all times.

Input Evaluation

The evaluation of input texts consists of two parts, the (MT) system-independent evaluation and the system-dependent evaluation. The MT-independent evaluation weighs the linguistic and terminological complexity of the text. The MT-dependent evaluation relates the complexities of the input text to the capabilities of the MT system that it will be subjected to.

Text Complexity Index (A)
The text complexity index describes the linguistic, terminological, and format complexity of input texts. Table 1 is an example list of the parameters that might be measured.

Table 1: Text Complexity Index

Grammatical Phenomena	❑
nested clauses, missing subjects, interrogative, etc.	(■ → etc.)
Terminological Domains	①
medical, space, machinery, weather terminology, etc.	(② ③ ④ ⑦ etc.)
Tables	
two columns, same width; five columns, variable width; etc.	➨ ➜ etc.)
Currencies	$
Japanese yen, British pounds, American cents, etc.	(¥ £ ¢ etc.)

The parameter category is represented through a symbol. Single parameters are represented through symbols inside parentheses. The number of parameters (and symbols) indicate the further stages of differentiation. The more differentiation symbols that are present, the more complex is the use of that parameter in the text. The more parameters present in the text, the more complex it is.

The list in Table 1 is just an example, the actual listing depends on the text for which a complexity index has to be calculated. The level of detail is selectable; it should be defined before the calculation starts, because it can not be changed afterwards.

The complexity index is an objective numerical characterization of a text. To a certain extent it will be possible to use parsers and equivalent software tools to objectively register some of the parameters. This index does not change, regardless of what procedure the text is subjected to, as long as no changes are made to the text. Assuming that MT output quality is proportional to MT input complexity, the complexity index allows for limited conclusions about the suitability of a text for MT.

Translatability Index (AB)
The MT-dependent evaluation of the text provides the text's translatability index as it pertains to the system for which it has been calculated. Since not all systems are designed to handle the same complexities (due to transfer modes or technical limitations) it is necessary to weigh those complexities in relation to the capabilities of the individual system. This feature is represented by the relation AB in Figure 1.

The translatability index compares the text complexity and the MT system's capabilities. In other words, in order to be able to make a statement about the translatability of a text by a specific MT system we need to know which of the text parameters (as given by the complexity index) the system can or is supposed to handle, according to its developer/vendor. Since it may not always be possible to obtain from vendors a list of the system's capabilities, a less accurate route may be taken: a test suite may be developed in order to determine features by abstraction. Such a test suite is inherently less accurate than a vendor-supplied list, since there will be no definite proof that it

actually measures the features it is designed to measure. As long as access to the actual machine processing information is denied, no test suite - no matter how intelligently designed - can claim to test with 100% accuracy what it is designed to test. There will always remain a degree of uncertainty, especially if the test results are negative, since the process might be derailed before it manages to test the parameter in question. Furthermore, in comparative evaluations the use of test suites introduces an extra dimension of complexity and error, since test suites are not necessarily transferable from one MT system to another.

Table 2 shows sets of complexity parameters and MT capabilities that are absolute matches.

Table 2. Translatability Index

Text Complexity	**MT Capabilities**
❑ (■ → etc.)	❑ (■ → etc.)
① (② ③ ④ ⑦ etc.)	① (② ③ ④ ⑦ etc.)
(➨ ➜ etc.)	(➨ ➜ etc.)
$ (¥ £ ¢ etc.)	$ (¥ £ ¢ etc.)

The table shows almost the highest possible rating for a translatability index, since the MT system has nearly all the capabilities necessary to handle all the selected complexities of the input text (it lacks only one of the terminology domains). This is certainly not a very likely scenario. In most cases there will be a more or less grave mismatch between text complexity and system capabilities. This shows how important it is to consider the relation between the input text and specific MT systems. Assuming that the degree of text translatability is proportionate to the quality of the output, it is not enough to determine the complexity of the text alone but it is also necessary to determine the overlap between text complexity parameters and system capability features.

MT System Evaluation

By weighing the text complexity in relation to the complexity handling capabilities of the MT system a system performance score for the specific system may be obtained. This performance index can be used to rank individual systems, and thus allow a comparison of different MT systems.

Performance Index (B)

The performance index can be obtained in two different ways: (1) in the form of a vendor/developer's description of a product's processing design and/or performance capabilities; or (2) in the form of evaluations of test suite translations. The vendor-supplied information is preferable not only because of cost and time considerations but also because of the possibility, already mentioned, that test suites might not measure the capability they are meant to measure - with, at best, unreliable results.

Since it measures capacities to handle text complexity, the performance index can be used to establish a ranking of MT systems. There is no advantage to a vendor/developer who overstates the capabilities of a system. If a vendor/developer makes claims about the handling of certain complexity parameters, and the evaluation shows that the feature is never handled correctly, then the system will be assigned a lower performance index within the group of MT systems. If a particular feature is not claimed by the vendor/developer, then the product will be assigned to a lower complexity index; however, because it is then found to handle the feature correctly it will be given a higher performance ranking. In general, a vendor/developer who does not overstate the capabilities of a system which receives a high performance rating should be considered more trustworthy than a vendor/developer who claims a capability to handle complexity higher than the system can actually achieve.

In practical applications, it is desirable to match text complexities and the capabilities of the MT system in order to optimize the translation process. If the MT system is not able to handle a number of the complexities contained in the source text, then output quality will presumably be less than adequate. On the other hand, on the assumption that MT system costs are proportionate to the range of a system's capabilities, it would not be cost-efficient to procure a system which can in fact handle more complex phenomena than actually required for the translation of source texts - even though output quality will presumably be adequate. Hence, it is in the best interests of both users and vendors/developers to state precisely the capabilities of a system and to seek to match text complexities to those capabilities.

MT systems (and the output they produce) are comparable only in so far as they possess the same capabilities. By determining the performance level for a system it (and its output) can then be compared to (that of) other comparable systems. If the output qualities have been compared (see following section on "Output Evaluation") and have been judged to be about equal, other factors such as the improvability index, user-friendliness, cost etc. should be taken into consideration when deciding the most suitable system for specific tasks.

Improvability Index (BC)

The relationship between an MT system and its output is a test of the validity of the vendor/developer's claims. More importantly, however, it permits some predictions about the improvability of the system.

If the MT capabilities match the complexities of input texts completely there is no need for system improvement. Few systems will achieve this, so it is important to find out whether what appears to be the most favored system can in fact be improved. A comparison between the performance index and the actual output will reveal the shortcomings of the system. Since these shortcomings can be determined with precision, the developer can then attend to any one of them and the evaluation can be repeated.

Ideally system performance should match the level of complexity. As said earlier, the level of detail of the complexity parameters will be defined by the evaluator based on the role envisioned for MT in the given environment. Detailed lists of complexities and of performance parameters will be available. A comparison of the (stated) performance features and the actual output provides the basis for specifying areas for further development to improve the system. Those issues that feature most prominently in the source text should be chosen as targets for improvements. Of course, the assessment of system improvability will be of interest only to those large volume customers willing to commit themselves to a long-term relationship with the MT vendor/developer. Likewise, developers will be willing to address specific developments only for such clients.

Output Evaluation

The output evaluation has two parts: evaluation of the quality of the text on its own; and evaluation of the quality of the text in relation to the input text.

Understandability Index (C)

The understandability index, just like the complexity index, is independent of the MT system. The level of detail of the parameters defining understandability is selectable, just as it is for the complexity index. The understandability index is a mixture of an objective numerical characterization of a text (grammatical parameters, terminological usage) and a subjective evaluation (content, natural language flow).

The parameters should include linguistic structures and terminological accuracy as well as content features. To a certain extent it will be possible to use parsers and equivalent software tools to objectively register the parameters pertaining to grammatical occurrences and perhaps also to provide lexical matching for terminological occurrences. In order to keep the content evaluation as objective as possible, I believe that it is necessary to study the degree of text understandability by various groups with various degrees of language skills and various degrees of domain knowledge.

The content evaluation should be performed by monolingual and bilingual readers. For each language group there should be readers knowledgeable about the domain and readers not knowledgeable about the domain. Part of the content evaluation should be an expression of opinion on the natural language flow. This factor could be substantiated by submitting the text to an appropriate grammar checker (i.e. one that knows the linguistic structures that are representative of this domain). The cumulative results will be a fair indication of the actual understandability of the output text.

Table 3. Understandability Index

	Linguistic Aspects	Terminology	Natural Lang. Flow	Content
Monolingual/No Domain Knowledge			➨	➨
Monolingual/Domain Knowledge		➨	➨	➨
Bilingual/No Domain Knowledge			➨	➨
Bilingual/Domain Knowledge		➨	➨	➨
Tools (Parser, Grammar Checker, Lexicon Matching, etc.)	➨	➨	➨	

Table 3 depicts the ideal situation where all linguistic aspects are evaluated by machine, i.e. a software tool. In reality this is not (yet) possible, because the necessary tools are not sophisticated enough (yet), and any human evaluator will be influenced in his/her assessment of one aspect by factors that belong to another aspect. For example, when evaluating language flow linguistic factors will influence decisions, and when evaluating content terminology will influence decisions. Where one aspect is assessed by more than one evaluator and/or by an evaluation tool there is the possibility of weighing the results according to factors such as objectivity, sophistication, etc.

Equivalency Index (AC)

The second component of the output evaluation is the comparison of input and output text. Historically, this has constituted the main part - and often the only part - of the evaluation of MT systems. However, since neither input texts nor MT systems are always "created equal" the differences have to be considered in every evaluation. If the differences are neglected and not incorporated in the evaluation, the method is subject to the accusation that it compares apples and oranges. An evaluation method that has a comparison as one of its core processes is only valid if it compares comparable things, factors, texts, systems, or whatever the object of comparison may be. A comparison of non comparable objects is flawed at best and meaningless in most cases.

The equivalency index is closely connected with the complexity and performance indices. If there is a match between the performance capability of the MT system and the complexity of the text, input and output should be equivalent. In those cases where there is no match or there is a mismatch between complexity and capability it is most likely that there is less or no equivalence. This assertion has to be vague because of the possibility of chance equivalence. Although there may be no match between complexity parameter and performance feature, the MT system might, for some reason or other, have produced output that is equivalent to the input.

The measurement of equivalency can be done - just as with indices of complexity, translatability, performance, and understandability - either at a very superficial or at a very deep level. The degrees of detail for all indices should be the same, in order to maintain the same overall depth of evaluation.

The equivalency measurement deals with errors in general (incl. spelling, grammatical and lexical errors), omissions and insertions, and also aspects of formatting in the broadest sense (i.e. indentations, tables, columns, dates, currencies etc.).

This equivalency checking procedure should ideally be done by software tools, since they are preferable to human checking by eliminating bias and factors of personal preference (cultural and social background, regional influences, etc.) Unfortunately in this area software tools are still underdeveloped. Parsers can help to "dissect" sentences and thus texts, but I know of no parsers that can check the equivalency of a source text and its translation. However, I would assume that it is possible. The same can be said for terminological/lexical matching programs.

I believe it is safe to assume that if the translatability index and the performance index match, the degree of equivalency will be high; and of course analogously, if the translatability index and the performance index do not match, the degree of equivalency will be low.

Conclusions

A differentiated text-related evaluation methodology such as the one proposed here allows predictions about a text's suitability for machine translation and about the likely degree of success if submitted to an MT system. Furthermore, it permits the ranking of machine translation systems and therefore a comparison of similar systems. Lastly, it provides information necessary for the matching of text requirements to appropriate MT systems.

The individual components of this type of evaluation are not self-contained but interrelated. Even though the complexity and understandability indices are MT-independent they still have an impact on the evaluation as a whole; viz. the complexity index, in that it provides a measurement of the source text on a scale of possible text complexity; and the understandability index, in that it provides an MT-independent assessment of the usefulness of the resulting translation. The combination of all components in one unified strategy offers complete and comparative evaluations of MT systems.

Figure 2 at the end of this paper shows the components and their interrelations. Whereas Figure 1 showed the components and relations in an order of succession from input text to output text, this figure shows the components and their interrelations with regard to their dependency on an MT system.

The indices connected in a triangle, i.e. the indices of Translatability, Performance and Equivalency, constitute the center of any evaluation. All three factors are MT-dependent. A high rating of any one of them will probably result in high ratings for both of the others; on the other hand, a low rating for any one of them will indicate that the ratings of the other two indices are likely also to be low.

The index to the right of the triangle, the Improvability Index, is a measure that is relative to the performance of the system. It is not only dependent upon the translation capabilities of the system and the complexities of the texts to be used with the system but also upon the physical constraints of the system. This index points outwards towards other features that are not text-related and that will thus have to be evaluated independently of this proposed methodology.

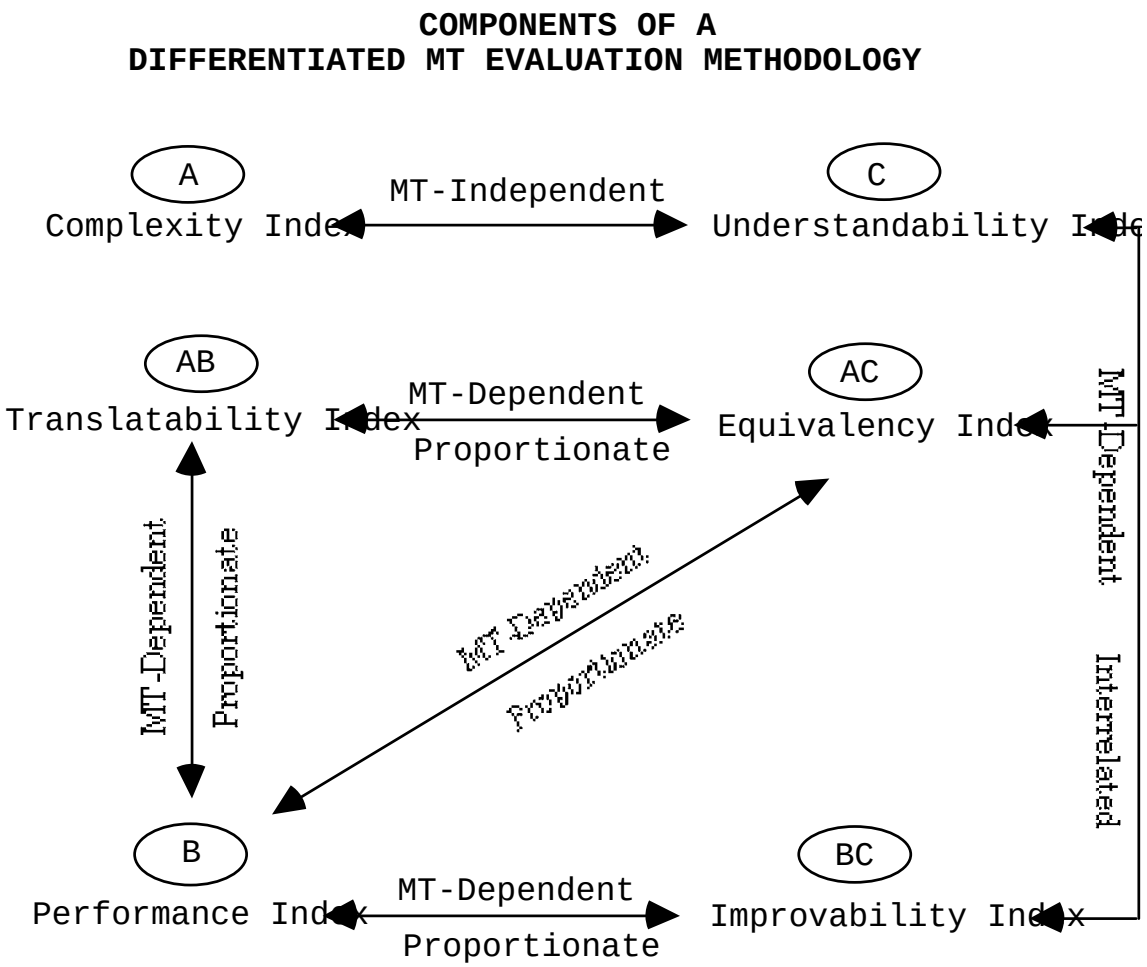

Figure 2: Components of a Differentiated MT Evaluation Methodology

READERS' FORUM

This section of MT News International is for readers to express personal views of issues 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.

A Marketing Machine

Colin Brace

With the recent announcement that Globalink and MicroTac intend to merge, the MT community is faced with a large and aggressive marketing-driven translation software vendor looming in its midst. Individually, the two companies, which together were good for more than US$10 million in sales last year, have both been surprisingly successful in selling translation packages to the non-language professional, that is, people who don't know a given foreign language and are susceptible to the prospect of acquiring some form of technological assistance for communicating in it. The combined entity, which will continue under the name of Globalink, will have a broad product line, from "entry-level" packages retailing for around fifty dollars to so-called "professional" versions, running under Unix and Windows, costing upwards of a thousand dollars.

For both established MT vendors as well as the R&D community, the rise of Globalink raises some significant issues. Long-time MT vendors will be confronted with a number such tough questions. What is the difference between this cheap MT system and that expensive one? How can users justify the extra cost? And, most important, how will developers maintain the differential between the two? The history of the IT industry suggests that low-end systems have a tendency to

improve more quickly than high-end ones, which, if it proves true in this field, could mean a few trusted names may soon find themselves in the lurch. The research community, meanwhile, may find these developments a mixed blessing. A field whose money-making potential is at long last "proven" will no doubt attract more funding, yet, ironically enough, with market prospects more clearly in view, there will be less interest -- and patience -- for tackling the "hard" problems.

Like it or not, today's existing and potential MT users will want "objective" benchmarks with which to compare the increasing number of systems now available. Not many will have, like the Union Bank of Switzerland, the personnel and resources for a thorough six-month evaluation of the potential candidates. While the topic of MT evaluation may seem like a bottomless black hole, it doesn't diminish the fact that easy to implement evaluation methodologies and reproducible quality metrics are now more needed urgently than ever before.

Multilingual MT and the Post-editing Bottleneck

Evgeny Lovtsky

In the account of PaTrans (MTNI#8) it is significant that Eurotra's failure is openly admitted. I have always been very sceptical about Eurotra's prospects of success. Multilingual systems in general seem to me an unrealistic proposition.

I would suggest that before embarking on developing ambitious multilingual MT project, the author should try his hand at a "simple" straightforward one-direction MT system for one language pair of his choice. If he succeeds at producing a life-size operational model (not just a prototype), he is welcome to add other languages and make further experiments. The experience he would acquire will save him much trouble in the future. It is my deep conviction that now efforts should be concentrated on eliminating the bottleneck of all commercial MT systems, namely *post-editing*. Post-editing is comparable in duration and complexity with translation itself. To solve the problem we must develop *an "intelligent" screen editor* that would take care of purely mechanical linguistic operations and leave to the human editor the intellectual part of the job. For example, it could supply synonymous translation equivalents (and thus make it possible for anyone *who may not know the source language* to edit machine translations.) It could substitute translations selected by the human editor for those chosen by the computer *at a click*, **making all necessary alterations in the sentence as it does so**. It could toggle, for example, between a participle attribute and an attributive clause *at a click of the mouse*. Or between perfective and non-perfective aspects of the Russian verb (to solve the problem of aspect in Russian algorithmically is practically impossible.) I can name dozens of linguistic operations which lend themselves to automation, saving time and effort of the human editor. Evidently, the most reasonable way is to make such an intelligent screen editor part of the MT system itself, so that it could use its dictionaries and other linguistic and software resources, as well as the internal representation of the translated text, produced by the MT system.

Evaluation of Integrated Translation Systems

Uwe Reinke

[From LINGUIST listserver, Vol-5-514.]

There is a recent tendency in machine-aided translation to develop systems that enable the translator to "recycle" former translation units. Besides a terminological database these memory-based translation systems - or integrated translation systems, as I prefer calling them - contain a database, the so-called translation memory, that stores translation units and compares source-language units currently to be translated to the units in the memory in order to find exact matches or "similar" sequences and make the former translation available again. This leads to the question whether and how translation units that are "similar" to a section currently to be translated are found in the memory. An even more interesting and fundamental question is "what does 'similarity' of translation sequences mean to the machine and what does it mean to a human being? And what kinds of

'similarities' are there in different types of texts?" These are some of the questions I would like to deal with in my PhD-thesis in order to find ways for the evaluation of integrated translation systems.

So far, I tried to analyse the linguistic performance of two commercial translation memory systems (IBM's Translation Manager/2 and Trados' Translator's Workbench II). As a first start, I tried to name some broad types of syntactic/semantic similarities, such as paradigmatic alterations, expansion of phrases by further attributes, altered position of phrases in a sentence, changed position of clauses, changes from passive voice to active voice etc. I used these types as categories for the analysis. The major results are contained in a paper that will be published by Langenscheidt in one of the next issues of *Lebende Sprachen*.

One major topic I would now like to focus on, is the aspect of "likeness" of (sentence) patterns in technical documentation. This would enable me to find a more detailed classification of semantic and syntactic "likeness" and to compare human understanding of this phenomenon to the performance of the different systems.

At the moment, my major problem is to collect a suitable amount of machine-readable texts in English, German and French. This is, why I would like to find out, whether there might be some list-members who could perhaps help solving this problem.

The kind of texts I am looking for could be roughly described as follows:

- "follow-up versions" of texts (i. e. different texts of the same text type belonging to the documentation of a newer and an older version of the same product)

- texts with different structures and/or functions belonging to one and the same product (e. g. online-help and user manual of a software)

- texts with similar functions belonging to different products of the same kind (e. g. a manual for a car produced by company A and a manual for a car produced by company B.)

I am basically interested in source-language texts. If I could also get hold of translations this would help testing the alignment tools now on the market, which are used to create translation memory databases from machine-readable source- and target-language texts.

I know that asking for text material - particularly for translations - might be quite a problematic thing. But as the material will be used for scientific purposes only, I think there may be some people around who might be able to contribute to my text corpus or might have some ideas where to get further assistance.

Please use my private e-mail address for replies: Uwe Reinke, Universität des Saarlandes, Fachrichtung 8.6, D-66041 Saarbrücken (Tel:+49/681/302-2929; Fax: +49/681/302-4440; Email: reinke@rz.uni-sb.de)

PAUL L. GARVIN (1919-1994) OBITUARY

Christine A. Montgomery
(Language Systems Inc.)

The recent death of Paul L(ucian) Garvin -- who pioneered machine translation research in the United States and was internationally known for his research on empirical methodology in linguistics --is untimely, for it has occurred just as linguists and computational linguists are in the process of rediscovering empirical methods and the role of data in linguistic research. In the changing intellectual climate in linguistics and computational linguistics, the significance of Garvin's contributions to these fields will be more widely recognized.

In 1952-4, Garvin was the principal scientist involved in the ground-breaking experimental collaboration between Georgetown University and IBM to produce the first public demonstration* of a machine translation system for Russian to English. From 1954 until 1959 at Georgetown, Garvin evolved the "fulcrum" approach to machine translation, in which the syntactic fulcrum or most highly information bearing constituent of a particular syntactic structure is identified and exploited to

analyze the given structure -- e.g., the fulcrum of a clause is the predicate, the fulcrum of a noun phrase is the head noun, etc. The basic translation strategy is to process the sentence by successively exploiting the lexical and syntactic information contained in the fulcrum (or fulcra) for a particular level of syntactic structures (e.g., noun phrases, prepositional phrases, participial structures), essentially developing a surface tree of the source language sentence via multiple passes through a sentence to identify and analyze fulcra of a particular type, going from bottom to top. Bar-Hillel's assessment of Garvin's approach (in a 1959 report to ONR on MT in the US and Great Britain) was that the "method seems to work rather satisfactorily for the syntactical analysis of a large class of Russian sentences, though its exact reach has not yet been fully determined nor all of its details debugged". Based on what Bar-Hillel had to say about other projects in that report, this was high praise indeed.

In 1960, Garvin joined Don Swanson at the Ramo-Wooldridge Corporation in Los Angeles, where he continued his work on the fulcrum approach to MT. Source language analysis was improved, and the generation of target language output was elaborated; it was created by an additional series of passes which provided appropriate inflections, reordered elements in phrases and clauses to achieve correct word sequences, and inserted articles based on lexical and syntactic features.

As in the empirically driven approaches which have recently resurfaced, the fulcrum approach to translation was corpus-based: corpora of scientific articles in a particular domain were "mined" by manual and automated techniques to identify lacunae in the lexicon and problem areas for syntactic analysis. In fact, looking at a list of projects under Garvin at Ramo-Wooldridge in 1962-3, which included two MT efforts, we also find 1) an automated text comparison development, used both for identifying new lexical items and for comparing machine to human translations to isolate problem areas; and 2) an automated lexicon development system for extracting specialized glossaries from bilingual dictionaries. Moreover, research carried out under the MT projects included statistical co-occurrence analyses of various types of fulcra with their associated arguments, e.g., predicates with assorted thematic roles, noun/adjective, preposition/noun, etc. These investigations presage the corpus-based and parallel-text CL studies of today. Another precursor of a current trend was a Garvin proposal to the National Science Foundation aimed at the collection of text corpora for use in MT and CL studies. Although the NSF response was favorable, that project -- along with most other MT and CL research --- was torpedoed by the famous ALPAC report of 1966.

After completing an NSF research project for computer simulation of informant work in linguistics (essentially an expert system for collecting and analyzing field data) in 1969, Garvin left industry and returned to academe, this time as Professor of Linguistics at the State University of New York at Buffalo. There, he immersed himself in sociolinguistic studies, focusing on the notion of standard language, with all the concepts and processes that implies, for diverse cultural contexts such as French Canada, Celtic areas of Europe and the British Isles, and indigenous cultures such as the Guarani of Paraguay.

After becoming a Professor Emeritus at Buffalo, in the last several years, Paul Garvin was able to divide his teaching and research activities between Buffalo and his native land of Czechoslovakia (now Czechia). Garvin took great delight in lecturing at Masaryk University in Brno, as well as at the University of Prague, and working with his Czech colleagues in both Brno and Prague. In 1990, he received an honorary doctorate from Masaryk University, recognizing his substantial contributions to Czech language and culture through the years (Garvin was responsible for translating the works of many of his Prague School mentors and colleagues, and for introducing their research in US and European linguistic circles).

Garvin was himself the student of Roman Jakobson, who introduced him to linguistics, and, according to Garvin, provided him with his professional orientation of functionalism in linguistics, based on Prague School theory. He also studied with the anthropological linguist, Carl Voegelin, at Indiana, gaining his lifelong interest in the study of unexplored languages and his expertise in linguistic field work.

Paul Garvin was an exceptional linguist in many ways. Unlike most of his linguist colleagues, he was also an accomplished polyglot, acquiring new languages with incredible ease. In informant work, Paul would simultaneously elicit utterances, analyze their structure, and begin generating novel grammatical utterances in the language under investigation, acquiring some facility with the language within a few work sessions.

He was also an exceptional human being -- a devoted teacher, a doting father, and a caring friend. We will not see his like again.

---

* For descriptions of the demonstration from the *New York Times* and the *Christian Science Monitor*, January, 1954, as well as insightful commentary from John Hutchins, see "From the Archives" in MTNI 8 (May 94): pp. 15-18.

PUBLICATIONS and DATABASES

Artificial Intelligence

Special issue devoted to Empirical Artificial Intelligence

Call for Papers, from the editors: Paul Cohen (cohen@cs.umass.edu) and Bruce Porter (porter@cs.utexas.edu).

We are looking for papers that characterize and explain the behaviors of systems in task environments. Papers should report results of studies of AI systems, or new techniques for studying systems. The studies should be empirical, by which we mean "based on observation" (not exclusively "experimental," and certainly not exclusively statistical hypothesis testing). Examples (some of which are already in the AI literature) include:

* A report of performance comparisons of message-understanding systems, explaining why some systems perform better than others in some task environments.
* A study of commonly-used benchmarks or test sets, explaining why a simple algorithm performs well on many of them
* A study of the empirical time and space complexity of an important algorithm or sample of algorithms
* Results of corpus-based machine-translation projects
* A paper that introduces a feature of a task that suggests why some task instances are easy and others difficult, and tests this claim
* Theoretical explanations (with appropriate empirical backing) of unexpected empirical results, such as constant-time performance on the million-queens problem
* A statistical procedure for comparing performance profiles such as learning curves
* A resampling method for confidence intervals for statistics computed from censored data (e.g., due to cutoffs on run times)
* A paper that postulates (on empirical or theoretical grounds) an equivalence class of systems that appeared superficially different, providing empirical evidence that, on some important measures, members of the class are more similar to each other than they are to nonmembers.

The empirical orientation will not preclude theoretical articles; it is often difficult to explain and generalize results without a theoretical framework. However, the overriding criterion for papers will be whether they attempt to characterize, compare, predict, explain and generalize what we observe when we run AI systems.

This is an atypical special issue because many of us think there is nothing special about empirical AI. It isn't a subfield or a particular topic, but rather a methodology that applies to many subfields and topics. We are concerned, however, that despite the scope of empirical AI, it might be underrepresented in the pages of the Artificial Intelligence Journal. This special issue is an experiment to find out: if the number of submitted, publishable papers is high, then we may conclude

that the Journal could publish a higher proportion of such papers in the future, and this issue might be inaugural rather than special.

Three principles will guide reviewers: Papers should be interesting, they should be convincing, and in most cases they should pose a question or make a claim. A paper might be unassailable from a methodological standpoint, but if it is an unmotivated empirical exercise (e.g., "I wonder, for no particular reason, which of these two algorithms is faster"), it won't be accepted. In the other corner, we can envision fascinating papers devoid of convincing evidence. Different interpretations of "convincing" are appropriate at different stages of projects and for different kinds of projects; for example, the standards for hypothesis testing are stricter than those for exploratory studies, and the standards for new empirical methods are of a different kind, pertaining to power and validity. If, however, the focus of a paper is a claim, then convincing evidence must be provided.

Deadline: Jan. 10, 1995. Please contact either of the editors as soon as possible to tell us whether you intend to submit a paper, and include a few lines describing the paper, so we can gauge the level of interest and the sorts of work we'll be receiving.

The Editorial Board for this issue includes: B. Chandrasekaran, Eugene Charniak, Mark Drummond, John Fox, Steve Hanks, Lynette Hirschman, Adele Howe, Rob Holte, Steve Minton, Jack Mostow, Martha Pollack, Ross Quinlan, David Waltz, Charles Weems.

New Journal of Translation and Interpreting

[From CORPORA Listserver]

The TRANSLATOR: Studies in Intercultural Communication. Devoted to bringing professional and academic interests closer together. Not restricted in scope to any particular school of thought or academic group. Managed by an editorial team with extensive academic and professional experience. Dedicated to: rigour, relevance, and readability.

Editor: Mona Baker (UMIST, UK). Editorial Board: Daniel Gile (ISIT, France); Ian Mason (Heriot-Watt, UK); Christiane Nord (Heidelberg, Germany); Anthony Pym (Spain); Lawrence Venuti (Temple University, USA); Judith Woodsworth (Concordia University, Canada). Review Editor: Myriam Salama-Carr (Salford, UK)

Two issues per year: 125 pages each. First Issue: April 1995.

First and second issues will include papers by: Lawrence Venuti (on copyright and authorship), Peter Fawcett (on translation and power play), Miriam Shlesinger (on cohesion in simultaneous interpreting), Keith Harvey (on compensation), Hannah Amit-Kochavi (on the translation of Israeli Arabic literature into Hebrew), Douglas Robinson (on women translators in the 16th & 17th centuries), Dirk Delabastita (on non-translation as a form of translation) and Candace Seguinot (on translating advertising texts).

Each issue will also include book reviews, including a review of an old but influential publication, plus a description of a translation/interpreting course. First issue: MA in Translation Studies (University of Surrey), Second Issue (Undergraduate courses at the Univeritat Autonoma de Barcelona).

Special Issue (Vol. 2, No. 2, 1996): Wordplay and Translation. Guest-Editor: Dirk Delabastita.

Order forms will be available by December 1994. To receive an order form, or if you would like a copy of the guidelines to contributors or reviewers, write to: The Translator, St. Jerome Publishing, 2 Maple Road West, Brooklands, Manchester, M23 9HH, UK. (Fax: UK 061-973-9856; Email: mona@ccl.umist.ac.uk)

Text Encoding Initiative Publishes Guidelines

In May, the Text Encoding Initiative (TEI) published its "Guidelines for Electronic Text Encoding and Interchange." This report is the product of several years' work by over a hundred experts in

fields ranging from computational linguistics to Ancient Greek literature. The Guidelines define a format in which electronic text materials can be stored on, or transmitted between, any kind of computer from a personal microcomputer to a university mainframe. The format is independent of the proprietary formats used by commercial software packages.

The TEI came into being as the result of the proliferation of mostly incompatible encoding formats, which was hampering cooperation and reuse of data among researchers and teachers. Creating good electronic texts is an expensive and time-consuming business. The object of the TEI was to ensure that such texts, once created, could continue to be useful even after the systems on which they were created had become obsolete. This requirement is a particularly important one in today's rapidly evolving computer industry.

To make them "future-proof", the TEI Guidelines use an international standard for text encoding known as SGML, the Standard Generalized Markup Language. SGML was originally developed by the publishing industry as a way of reducing the costs of typesetting and reuse of electronic manuscripts but has since become widely used by software developers, publishers, and government agencies. It is one of the enabling technologies which will help the new Digital Libraries take shape.

The TEI Guidelines go beyond many other SGML applications currently in use. Because they aim to serve the needs of researchers as well as teachers and students, they have a particularly ambitious set of goals. They must be both easily extensible and easily simplified. And their aim is to specify methods capable of dealing with all kinds of texts, in all languages and writing systems, from any period in history.

Consequently, the TEI Guidelines provide recommendations not only for the encoding of prose texts, but also for verse, drama, and other performance texts, transcripts of spoken material for linguistic research, dictionaries, and terminological data banks.

The Guidelines provide detailed specifications for the documentation of electronic materials, their sources, and their encoding. These specifications will enable future librarians to catalogue electronic texts as efficiently and reliably as they currently catalogue printed texts.

The TEI Guidelines also provide optional facilities which can be added to the set of basic recommendations. These include methods for encoding hypertext links, transcribing primary sources (especially manuscripts), representing text-critical apparatus, analyzing names and dates, representing figures, formulae, tables, and graphics, and categorizing of texts for corpus-linguistic study. The Guidelines also define methods of providing linguistic, literary, or historical analysis and commentary on a text and documenting areas of uncertainty or ambiguity.

The TEI Guidelines have been prepared over a six-year period with grant support from the U.S. National Endowment for the Humanities, Directorate General XIII of the Commission of the European Union, the Andrew W. Mellon Foundation, and the Social Science and Humanities Research Council of Canada. The effort is largely the product of the volunteer work of over a hundred researchers who donated time to share their experience in using computers and to work out the specific recommendations in the Guidelines.

The project is sponsored by three professional societies active in the area of computer applications to text-based research: the Association for Computers and the Humanities, the Association for Literary and Linguistic Computing, and the Association for Computational Linguistics, which have a combined membership of thousands of scholars and researchers worldwide.

Many projects in North America and Europe have already declared their intention of applying the TEI Guidelines in the creation of the large scale electronic textual resources which are increasingly dominating the world of humanities scholarship.

TEI P3, the Guidelines for Electronic Text Encoding and Interchange, is available in the following forms:

- in paper (1300 pp., 2 volumes), at a cost of $75 US, 50 pounds sterling, or 7500 yen. Order form below.

- electronically in an SGML-tagged form (ca. 5.6 Mb) using the TEI DTD documented in TEI P3, with minor extensions; this form is available without cost via Listserv or anonymous ftp. More info below.

- electronically in a formatted 'ASCII-only' version (ca. 3.1 Mb) suitable for display by those without an SGML-aware rendering engine; this form is available without cost via Listserv or anonymous ftp.

The TEI document type definition (DTD) files are available electronically via Listserv or anonymous ftp.

The electronic forms of the documentation are available via Listserv commands from LISTSERV@UICVM.UIC.EDU, and by anonymous ftp from:

ftp-tei.uic.edu (in pub/tei and its subdirectories)
sgml1.ex.ac.uk (in tei/p3 and its subdirectories)
TEI.IPC.Chiba-u.ac.jp (in /TEI/P3)
ftp.ifi.uio.no (in pub/SGML/TEI)

To fetch TEI P3 from the TEI-L file server maintained at the University of Illinois at Chicago, send electronic mail to

listserv@uicvm.uic.edu

containing one or more of the following lines.
To order the SGML-tagged version of TEI P3, include the line

get teip3 package

To order the formatted, untagged (ASCII-only) version of TEI P3, include

get p3ascii package

To order the TEI P3 DTD files, include

get p3dtds package

If you want ALL THREE packages (SGML-tagged, formatted, and DTDs), you may include all three of the lines given above, or the single line

get p3all package

For further information on using the file server, include the line

get edj8 memo

or consult the materials Listserv sent you when you subscribed to TEI-L. If you are not already subscribed, you can subscribe by including the following line in your note to Listserv@uicvm.uic.edu:

subscribe tei-l J. Doe

(substituting your name for 'J. Doe')

To get paper copies of TEI P3, contact one of the following for an order form: C. M. Sperberg McQueen, University of Illinois at Chicago, Academic Computing Center (M/C 135), 1940 W. Taylor, Rm. 124, Chicago IL 60612-7352, U.S.A.; TEI Orders, Oxford University Computing Services, 13 Banbury Road, Oxford OX2 6NN; Prof. Syun Tutiya, Faculty of Letters, Chiba University, 1-33 Yayoi-cho, Inage-ka, Chiba 263 Japan (Fax: +81 (43) 256-7032; Email: tutiya@culle.l.chiba-u.ac.jp).

Publication on Testing Morphological Analyzers

[From ELSNET listserver]

The first Morpholympics, held March 7 and 8, 1994, at the University of Erlangen-Nuremburg, was organized as a competition between different systems of automatic word form analysis, testing for coverage, speed, and quality of analysis.

A detailed description of the Morpholympics has now been published in the latest issue of the *LDV-Forum* (Vol. 11:1), containing:

1. The judges' evaluation of participating systems (in German)
2. Presentation of the Kimmo-system (in English)
3. Presentation of the Morph-system (in German)
4. Presentation of the LA-Morph system (in German)
5. Summary of the measurements (in English)
6. The coordinator's final report (in English)

Copies of the current LDV-Forum may be obtained by sending email to:

ute@iaisun.iai.uni-sb.de (There is a charge of DM 20.- plus postage and handling.)

The LDV-Forum is published by the Gesellschaft für Linguistische Datenverarbeitung (GLDV) and distributed to its members. If you would like to join GLDV, please send email to:

lenders@uni-bonn.de

A complete description of the First Morpholympics, including the presentations of all participating systems, is currently in preparation.

LRE Activity Report available

[From *I&T Magazine & News Review* (CEC DGXIII), Spring 1994]

The above report outlines DG XIII/E's activities in the Linguistic Research and Engineering (LRE) sub-programme within the Telematics programme. The Community has been investing in language processing for almost 20 years. The Commission entered the machine translation domain working on the Systran system in 1975. The Eurotra programme was funded to build a prototype machine translation system for the nine official Community languages. Within Esprit, the Community funded a number of speech and written language projects. More recently, the LRE sub-programme was operational from 1991-1994, and currently the Commission is laying the foundation for a follow-up to these activities within the forthcoming Fourth Framework programme.

In the LRE sub-programme the key areas for action identified in two calls for proposals launched in 1991 and 1992 were: general research; common resources, tools and methods to build a comprehensive linguistic infrastructure; and pilot applications to demonstrate the integration of language engineering technologies and components within information and communication systems. In total, 26 projects and accompanying actions were selected for funding. A total of 160 companies and organisations from throughout Europe are participating, with a significant presence of user bodies, SMEs and multinational companies. The tools, resources and pilot applications emanating from these projects will serve as an invaluable foundation for further exploration and exploitation by European academia and industry in a technology area which is vital to Europe's future.

The activity report, accompanied by project synopses, is available from: Robert Cencioni, CEC, DGXIII/E/4, L-2920 Luxembourg (Tel: +352 4301 32886; Fax: + 352 4301 34999.)

News from ISIR

[From ELSNET listserver]

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 free of charge via e-mail. If you wish to receive our publication, please send your e-mail address to flash@citi.doc.ca.

Online Database
ISIR's online database on information technology is also accessible via Internet.

Searching instructions are as follow:

1. Telnet SIRI.CITI.DOC.CA or Telnet 142.62.1.7
2. After the prompt ":", enter

HELLO identification,SIRI.CITI

For "identification", enter your name or the name of your organization (maximum length: 8 characters).

Example: HELLO Alex,SIRI.CITI

If you have any problems, please contact us at siri@citi.doc.ca.

LFG List

Mary Dalrymple

[From ELSNET Listserver]

I'd like to announce the formation of a new mailing list. The LFG List is a forum for discussion of linguistic, mathematical, and computational issues within the framework of Lexical-Functional Grammar, including:

- linguistic observations and their treatment in LFG
- formal devices and their `fit' to linguistic data
- computational implementation of LFG

To subscribe to the LFG List, send a message to

majordomo@list.stanford.edu

containing only the following line:

subscribe lfg

You will get back a message welcoming you to the list. To make a submission to the LFG List, send it to:

lfg@list.stanford.edu

We look forward to your participation!

Mary Dalrymple, Information Sciences and Technologies Laboratory, Xerox PARC.

CMU Artificial Intelligence Repository
and Prime Time Freeware for AI

[From LINGUIST list]

The CMU Artificial Intelligence Repository was established by Carnegie Mellon University to contain public domain and freely distributable software, publications, and other materials of interest to AI researchers, educators, students, and practitioners. The AI Repository currently contains more than a gigabyte of material and is growing steadily.

The AI Repository is accessible for free by anonymous FTP, AFS, and WWW. A selection of materials from the AI Repository is also being published on CD-ROM by Prime Time Freeware and should be available for purchase at AAAI-94 or direct by mail or fax from Prime Time Freeware (see below).

Accessing the AI Repository.

To access the AI Repository by anonymous FTP, ftp to:

ftp.cs.cmu.edu [128.2.206.173]

and cd to the directory:

/user/ai/

Use username "anonymous" (without the quotes) and type your email address (in the form "user@host") as the password.

To access the AI Repository by AFS (Andrew File System), use the directory:

/afs/cs.cmu.edu/project/ai-repository/ai/

To access the AI Repository by WWW, use the URL:

http://www.cs.cmu.edu:8001/Web/Groups/AI/html/repository.html

Be sure to read the files 0.doc and readme.txt in this directory.

Contents of the AI Repository

The AI Programming Languages and the AI Software Packages sections of the repository are "complete". These can be accessed in the lang/ and areas/ subdirectories of the AI Repository. Compression and archiving utilities may be found in the util/ subdirectory. Other directories, which are in varying states of completion, are events/ (Calendar of Events, Conference Calls) and pubs/ (Publications, including technical reports, books, mail/news archives).

The AI Programming Languages section includes directories for Common Lisp, Prolog, Scheme, Smalltalk, and other AI-related programming languages.

The AI Software Packages section includes subdirectories for:

agents/ Intelligent Agent Architectures
alife/ Artificial Life and Complex Adaptive Systems
anneal/ Simulated Annealing
blackbrd/ Blackboard Architectures
bookcode/ Code From AI Textbooks
ca/ Cellular Automata
classics/ Classical AI Programs
constrnt/ Constraint Processing
dai/ Distributed AI
discover/ Discovery and Data-Mining
doc/ Documentation
edu/ Educational Tools
expert/ Expert Systems/Production Systems
faq/ Frequently Asked Questions
fuzzy/ Fuzzy Logic
games/ Game Playing
genetic/ Genetic Algorithms, Genetic Programming, Evolutionary Programming
icot/ ICOT Free Software
kr/ Knowledge Representation, Semantic Nets, Frames, ...
learning/ Machine Learning
misc/ Miscellaneous AI
music/ Music
neural/ Neural Networks, Connectionist Systems, Neural Systems
nlp/ Natural Language Processing (Natural Language Understanding, Natural Language Generation, Parsing, Morphology, Machine Translation)
planning/ Planning, Plan Recognition
reasonng/ Reasoning (Analogical Reasoning, Case Based Reasoning, Defeasible Reasoning, Legal Reasoning, Medical Reasoning, Probabilistic Reasoning, Qualitative Reasoning, Temporal Reasoning, Theorem Proving/Automated Reasoning, Truth Maintenance)
robotics/ Robotics
search/ Search
speech/ Speech Recognition and Synthesis
testbeds/ Planning/Agent Testbeds
vision/ Computer Vision

The repository has standardized on using 'tar' for producing archives of files and 'gzip' for compression.

Keyword Searching of the Repository:
To search the keyword index by mail, send a message to:

ai+query@cs.cmu.edu

with one or more lines containing calls to the keys command, such as:

keys lisp iteration

in the message body. You'll get a response by return mail. Do not include anything else in the Subject line of the message or in the message body. For help on the query mail server, include:

help

instead.
A Mosaic interface to the keyword searching program is in the works. We also plan to make the source code (including indexes) to this program available, as soon as it is stable.

Contributing Materials to the Repository:

Contributions of software and other materials are always welcome, but must be accompanied by an unambiguous copyright statement that grants permission for free use, copying, and distribution, such as:

- a declaration that the materials are in the public domain,

or

- a copyright notice that states that the materials are subject to the GNU General Public License (cite version),

or

- some other copyright notice (we will tell you if the copying permissions are too restrictive for us to include the materials in the repository)

Inclusion of materials in the repository does not modify their copyright status in any way.
Materials may be placed in:

ftp.cs.cmu.edu:/user/ai/new/

When you put anything in this directory, please send mail to

ai+contrib@cs.cmu.edu giving us permission to distribute the files, and state whether this permission is just for the AI Repository, or also includes publication on the CD-ROM version (Prime Time Freeware for AI).

We would appreciate if you would include a 0.doc file for your package; see /user/ai/new/package.doc for a template. (If you don't have the time to write your own, we can write it for you based on the information in your package.)

Prime Time Freeware for AI (CD-ROM):

A portion of the contents of the repository is published annually by Prime Time Freeware. The first issue consists of two ISO-9660 CD-ROMs bound into a 224-page book. Each CD-ROM contains approximately 600 megabytes of gzipped archives (more than 2 gigabytes uncompressed and unpacked). Prime Time Freeware for AI is particularly useful for folks who do not have FTP access, but may also be useful as a way of saving disk space and avoiding annoying FTP searches and retrievals.

Prime Time Freeware helped establish the CMU AI Repository, and sales of Prime Time Freeware for AI will continue to help support the maintenance and expansion of the repository. It sells (list) for US$60 plus applicable sales tax and shipping and handling charges. Payable through Visa, MasterCard, postal money orders in US funds, and checks in US funds drawn on a US bank. For further information on Prime Time Freeware for AI and other Prime Time Freeware products, please contact: Prime Time Freeware, 370 Altair Way, Suite 150, Sunnyvale, CA 94086 USA (Tel: +1 408-433-9662; Fax: +1 408-433-0727; Email: ptf@cfcl.com)

Repository Maintainer:

The AI Repository was established by Mark Kantrowitz in 1993 as an outgrowth of the Lisp Utilities Repository (established 1990) and his work on the FAQ (Frequently Asked Questions)

postings for the AI, Lisp, Scheme, and Prolog newsgroups. The Lisp Utilities Repository has been merged into the AI Repository. Bug reports, comments, questions and suggestions concerning the repository should be sent to Mark Kantrowitz <AI.Repository@cs.cmu.edu>. Bug reports, comments, questions and suggestions concerning a particular software package should be sent to the address indicated by the author.

Release of European Corpus Initiative CD-ROM

[From ELSNET and CORPORA listservers]
The European Corpus Initiative is pleased to announce the availability of their Multilingual Corpus 1 (ECI/MCI) on CD-ROM.

Overview of the contents of ECI/MCI

The ECI/MCI corpus contains almost 100 million words in 27 (mainly European) languages. It consists of 48 opportunistically collected component corpora (many thanks to those who generously donated material!) marked up using TEI P2 conformant SGML (to varying levels of detail), with easy access to the source text without markup. 12 of the component corpora are multilingual parallel corpora with from two to nine sub-corpora. All the alphabetic corpora (there is some Japanese and Chinese) are encoded in the ISO LATIN family of 8-bit character sets (ISO 8859-1, -5 and -7). The CD-ROM is in High Sierra format (ISO 9660), readable on UN*X, MSDOS and Apple systems at least.

The component corpora vary considerably in size -- some of the larger ones:

ger03: German Newspaper texts from the Frankfurter Rundschau July 1992 - March 1993. Provided by Universität Gesamthochschule Paderborn Germany. Approximately 34 million words.

fre01: French Newspaper texts from Le Monde September, October 1989, and January 1990. Provided by LIMSI CNRS, France. Approximately 4.1 million words

dut02: Extracts from the Leiden Corpus of Dutch (newspapers, transcribed speech, etc). Provided by Instituut voor Nederlandse Lexicologie, Leiden, Netherlands. Approximately 5.5 million words

mul05: International Labour Organisation reports of the Committee on Freedom of Association 1984-1989. Parallel texts in English, French and Spanish. Approximately 1.7 million words per language

How to Acquire the ECI/MCI CD-ROM

The CD-ROM is available in the US from the Linguistic Data Consortium (LDC), for members of the LDC or those making a bulk purchase, and otherwise from ELSNET, 2 Buccleuch Place, Edinburgh EH8 9LW, SCOTLAND. The cost from ELSNET is 20 UK pounds plus postage, handling and tax where applicable, on signature of the necessary User Licence Agreements.

The ordering procedure is described in detail in
http://www.cogsci.ed.ac.uk/elsnet/eci.html
For those unable to access this, blank licence agreements can be retrieved via anonymous ftp from
ftp.cogsci.ed.ac.uk:pub/elsnet/eci/user-license.tex or .ps
Further information about ordering etc. should be requested from
elsnet@cogsci.ed.ac.uk
A complete contents listing can be accessed via the above ELSNET URL, or by anonymous ftp from
ftp.cogsci.ed.ac.uk:pub/elsnet/eci/mci-listing

Dutch Corpus on-line

[From CORPORA listserver]

The Institute for Dutch Lexicology INL (Instituut voor Nederlandse Lexicologie) offers you the possibility to consult a text corpus of ca. 5 million words of present-day Dutch text, by the international computer network. This corpus is different from the Dutch INL corpora on the ECI/MCI CD-ROM distributed by the Linguistic Data Consortium and ELSNET. The texts are derived from books, magazines, newspapers and TV broadcasts, and cover several topics such as journalism, politics, environment, linguistics, leisure and business & employment. You can easily define subcorpora on the basis of these parameters.

The retrieval system allows you to search for single words or for word patterns, including some predefined syntactic patterns that can be changed by the user. Searches concern the levels of word form, part of speech (POS), and head word, both separately and in combination by use of Boolean operators and proximity searches. During the search, data concerning frequency and distribution over the texts are provided at several levels. The output most often is a list of items, or a series of concordances (words in context) with a variable, user-defined textual context. Sorting facilities may make your analysis of the output data more comfortable. With some limitations due to copyright, the output of your searches can be transfered to your own computer by e-mail. It is not allowed to transfer complete texts or substantial text parts.

Most of the data has not been corrected, neither on the level of the text, nor on the level of POS and headword. POS and headword have automatically been assigned to the word forms in the electronic text by lingware developed at the INL.

The providers of the texts have given permission for use of their materials for non-commercial, research purposes only. The conditions for commercial use are still topic of discussion.

In order to get access to this corpus, an individual user agreement has to be signed. An electronic user agreement form can be obtained from our mailserver Mailserv@Rulxho.Leidenuniv.NL. Type in the body of your e-mail message: SEND [5MLN94] AGREEMNT.USE. Please make a hard copy of the agreement form, sign it, and return the signed copy to: Institute for Dutch Lexicology INL P.O. Box 9515, 2300 RA Leiden fax: 31 71 27 2115

If you need additional information, please send an e-mail message to Helpdesk5mln@Rulxho.Lei- denuniv.NL, or send a fax to Mrs. dr. J.G. Kruyt.

SUSANNE Corpus: Release 3 Available

Geoffrey Sampson

[From ELSNET listserver]

Release 3 of the SUSANNE Corpus is now complete and is available, like earlier releases, by anonymous ftp from the Oxford Text Archive. Release 3 incorporates several thousand modifications dealing with errors and inconsistencies in the Corpus which came to light during the process of preparing the book ENGLISH FOR THE COMPUTER for publication. It also includes additional information in the documentation file.

To obtain a copy of SUSANNE Release 3, log in by anonymous ftp to: black.ox.ac.uk,
move to the directory: ota/susanne,
and follow the instructions in the README file in that directory.

A number of users have enquired about the publication schedule for the book. The manuscript of ‘English for the Computer’ was delivered to Oxford University Press in August 1993, and publication is expected late in 1994.

For those not familiar with the SUSANNE Corpus: this is an annotated sample comprising about 130,000 words of written American English text, produced to exemplify a set of annotation standards which attempt to specify an explicit notation for all aspects of the surface and logical grammar of real-life English in sufficient detail that analysts independently applying the standards to the same text must produce identical annotations. These standards are defined in the book `English for the Computer'; a skeleton outline of the scheme is included in the electronic documentation file

which accompanies the Corpus. The texts of the SUSANNE Corpus are a subset of the texts included in the (unannotated) Brown University Corpus.

"Terminology" - a new journal

The importance of terminology for translation, documentation, classification and scientific communication has long been recognised. Its impact in MT work has increased significantly in recent years with the development of domain-specific and sublanguage systems and with the widespread use of controlled vocabulary for input texts. The broader perspectives are sketched in their introductory essays by the two editors, Helmi B.Sonneveld (Amsterdam) and Kurt L.Loening (Columbus, Ohio), who were responsible last year for a collection, also published by Bejamins, which may be regarded as a precursor of this new journal (*Terminology: applications in interdisciplinary communication* [see MTNI#6]). The wider issues are also explored by Juan Sager, a doyen of translation and terminology - author of *A practical course in terminology processing* (Benjamins, 1990) and recently a major monograph on translation and its automation (*Language engineering and translation: consequences of automation* [see 'Publications Received']). Each of these authors in their own ways emphasise the significance of terminology in facilitating and enhancing the communication and interchange of specialised knowledge and information within and across languages; and to underline this central role of terminology studies the journal has been given the subtitle: International Journal of Theoretical and Applied Issues in Specialized Communication.

Other articles in this first issue discuss theoretical issues of terminology, the current standards for term description, the terminology of acupuncture and telecommunications, comparison of Japanese and English term structures, terminology management in machine translation, and automatic term recognition. [For details see 'Publications Received'.] In addition there are survey articles, short notices, an obituary (Wayne P.Ellis), book reviews, reports of meetings, and a calendar of conferences. The high quality of this issue reinforce the welcome which should be given to this new journal in an increasingly important field.

Terminology and MT

The book published by Jörg Schütz (*Terminological knowledge in multilingual language processing*) in the series "Studies in Machine Translation" from the European Commission [for details see 'Publications Received'] presents a model of an MT system in which terminological knowledge plays a central role. The most important domain-specific knowledge in technical translation is that which is embodied in the terminology of the domain. The author illustrates how the encoding of definitions for technical terms can be employed directly in an MT system. The author develops a new approach which utilizes terminological knowledge within a model of processing based on the unification grammar formalism ALEP. The basic linguistic knowledge is represented in a competence grammar; the subject domain knowledge is encoded in a hierarchy of typed feature terms derived from terminological sources, and this specialised knowledge constrains analysis and transfer. For an illustrative subject domain the author has chosen the area of satellite communications. Though small in compass, the demonstration merits attention as an example of an innovative approach to multilingual MT.

Proposals for MT evaluation

In a discussion report presented to the Commission of the European Communities, Antoni G. Torrens (*MT evaluation and quality benchmarks* [for details see 'Publications Received']) has outlined the criteria and parameters which he believes should be adopted in evaluations of MT systems. He observes that there is very little consensus ("Doubt, hesitation, confusion and chaos seem to prevail") and that MT evaluators have ignored the lessons which can be drawn from

language testing experts. He comments that "language testing covers all forms of linguistic competence and performance, so it can cover translation as well" and, since it covers translation testing, it is also applicable to MT evaluation. The essence of the author's proposals is that an operational MT system for a given language pair should be tested as if it were a student taking an exam. "First, you give the student or the MT pair a performance test (a comprehensive battery on translation proficiency)". However, if he/she or it fail, you may go on to look within the black box with "diagnostic tests" and 'glass box' approaches where designers and developers give explanations and evaluators perform subtests in order to reach personal assessments. These cannot have the stringency of scientific tests but are the only option if some kind of evaluation has to be done on MT prototypes and whenever financial investment has to be decided.

The author proposes three tests: a) "for multi-purpose comparable quality evaluation of a translation sample or corpus"; b) "for purchasing and similar decisions involving comprehensive study of technically and economically relevant features"; and c) "for regular checks of development progress of any given pair of operational MT systems". For the second he recommends following the advice given in earlier works (e.g. the JEIDA proposals [see MTNI#4: 19-20]). He makes explicit suggestions only for the first type of test: a *standardized translation proficiency test*, which should be based on the pattern of foreign language proficiency tests. The aim should be a test with high levels of reliability (absence of extraneous factors, significant sampling, intermarker reliability), with appropriate validity (e.g. in terms of communicative efficiency), with accepted norms (in order to reduce subjectivity), and with reasonable practicality (i.e. not costing too much to apply). As a first step to developing such a proficiency test, the author recommends the formation of a permanent evaluation board, commissioned with the production of a complete specification. The board should embrace expertise in "general and applied linguistics, translation theory and practice, translation teaching and marking, educational statistics and foreign language testing, computational linguistics and MT, and functional communication and translation use." The range of expertise illustrates the author's conviction of the complexity of MT evaluation and the importance of achieving substantial acceptability of any evaluation method the board may propose. The next step would be for the board to design and develop the "translation proficiency testing instruments needed for intersystem and interpair comparisons." The final task of the board would be to make proposals for the other tests, (b) and (c) above. The discussion paper is accompanied by an annotated list of references in relevant areas and an appendix drafting an evaluation methodology based on the principles described in the body of the text. In sum, this is a document which ought to be examined in detail by any group developing evaluation methods. It is not available for sale, but a shorter version will be published in the journal *Terminologie et Traduction*.

Eurotra Final Review Report

[From *I&T Magazine (CEC)*, Spring 1994]

Eurotra is a Community R&D programme which was initially approved by the Council in November 1982 and terminated at the end of 1992. Its objective was to create a prototype MT system of advanced design capable of dealing with all official EC languages. The programme was carried out in the framework of Contracts of Association concluded with all Member States and co-financed by the EC and Member States. The total Community budget over the 10 years was 37.5 million ecu.

The final evaluation of the programme was undertaken in 1993 by a panel of high-level personalities, chaired by Dr Brian Oakley, former director of the Alvey programme. Although the Commission did not entirely succeed in creating an operational multilingual MT system prototype capable of handling all official EC languages, the conclusions of the report are nevertheless very positive. It stresses in particular the role Eurotra has played in educating a generation of computational linguists, in shifting the centre of gravity of computational linguistics research from the West Coast of the USA to Europe, and the influence it had in devising industry-led projects like Eurolang and both Community and national research programmes.

It also formulates recommendations for a broad technology approach for the future. These recommendations have largely been incorporated in the Community's proposal for a language engineering programme as part of the Telematics specific programme of the Fourth Framework programme.

The Eurotra Final Review Panel Report is available from the Commission services. Contact: Sergei Perschke, CEC, DGXIII/E/5, L-2920 Luxembourg. (Tel: +352 4301 33423; Fax: +352 4301 34655.)

PUBLICATIONS RECEIVED

Journals

BCS (British Computer Society) Natural Language Translation Specialist Group Newsletter *22 (April 1994)*. Contents include: p.6-8: Standards [on EAGLES, ELSNET, TEI, RELATOR]. -- p.10-12: Translator's Workbench: multilingual documentation and communication (Khurshid Ahmad). -- p.12-13: The evaluation of machine translation (Siety Meijer). -- p.13-15: ATAMIRI computer translation system (David Stanton). -- p.15-17: The Alvey natural language tools (John Carroll). -- p.17-20: The MicroCat machine assisted translation system: an assessment project at the University of Exeter (Derek Lewis). -- p.20-23: Hugo Plus - a brief review (Glyn Phillips). -- p.23-24: AECMA Simplified English (David Wigg).

Computational Linguistics, *vol.20, no.2 (June 1994)*. p.155-172: Tagging English text with a probabilistic model (Bernard Merialdo). -- p.173-191: Tree-adjoining grammar parsing and Boolean matrix multiplication (Giorgio Satta). -- p.193-232: Japanese discourse and the process of centering (Marilyn Walker, Masayo Iida, and Sharon Cote). -- p.233-287: Tracking point of view in narrative (Janyce M.Wiebe). -- p.289-300: Parsing and empty nodes (Mark Johnson and Martin Kay). -- p.301-317: RAFT/RAPR and centering: a comparison and discussion of problems related to processing complex sentences (Linda Z.Suri and Kathleen McCoy).

Elsnews, *vol.3 no.2 (June 1994)*. Contents include: p.2: An eyewitness account of the HCM Workshop on Dialogue and Discourse (Paul McKevitt). -- p.4-5: ALEP, arriving at the next platform (Paul Meylemans). -- p.5-7: Reusability of grammatical resources (Gregor Erbach, et al.). -- p.9: ELU, the environment of choice for Macintosh computers (Dominique Estival, et al.)

INL Infoterm Newsletter *70 (December 1993)*. Workshop on "Teaching methods and teaching material for terminolgy", Cologne, 23 August 1993. -- Joint Inter-Agency meeting on computer-assisted teminology and translation, Geneva, 30-31 August 1993. -- Visit of the Assistant Director General for Communication, Information and Informatics of UNESCO, Mr.Yushkiavitshus, to Infoterm. --Tekom Conference, Fredrichshafen, 4-5 November 1993. -- UNESCO General Conference in Paris, 25 October - 15 November 1993. --Meeting on quality and terminology work, Cologne, 8 November 1993. -- WHOTERM expansion. -- Chinese termbank at the Chinese Publishing House, Beijing, China.

Language Industry Monitor *no.21 (May-June 1994)*. Contents include: p.1-6: The Finnish formula [Kielikone MT system]. --p.7: LISA: back for more [LISA meeting in Heidelberg]. -- p.8-10: MT systems from Russia: a hot find? [marketing of SILOD by Dutch company]. -- p.9: Stylus: also from St Petersburg. -- p.10-11: Sietec's Metal (also) does Russian (Klaus Schubert) [extract in this issue]. -- p.13-14: Logos: working on front ends.

Language International, *vol.6 no.3 (June 1994)*. Contents include: p.7-8: Smart release: MaxTrans. -- p.13-14: Enter Eurolang [on the Optimizer]. -- p.15: Systran and Xerox announce long-term agreement. *vol.6 no.4 (August 1994)*. Contents include: p.3-6: TRANSIT for Windows (Sabine Nixon). -- p.7: XL8 verges on excellence (Tony Roder).

LISA Forum Newsletter, *vol.3 no.2 (May 1994)*. p.1: New challenges in localization (Colin Brace). -- p.3-6: Problems, trends and facts for the localization industry or, Stress reduction is our real motivation! (Jan Pfefferkorn). -- p.6-7: COMPAQ Inc. reorganizes their localization and internationalization procedures. -- p.8: OSF initiates localization program for DCE. -- p.10-12: What are the GLOSSASOFT deliverables? -- p.13-26:

Book reviews, by Chiaki Ishikawa and by Bob Peterson, of Ken Lunde 'Understanding Japanese information processing'.

Literary & Linguistic Computing, *vol.9 no.1, 1994*. Contents include: p.29-46: Corpora and computational lexica: integration of different methodologies of lexical knowledge acquisition (R.Bindi et al.). -- p.47-54: A generic model for reuseable lexicons: the Genelex project (M.H.Antoni-Lay et al.). -- p.55-64: Use and importance of standard in electronic dictionaries: the compilation approach for lexical resources (H.Khatchadourian et al.). -- p.65-78: Text processing using multilingual resources at the Computing Research Laboratory (J.Cowie et al.). -- p.79-86: Spoken language assessment in the European context (A.Fourcin and D.Gibbon). *vol.9 no.2, 1994*. Contents include: p.137-148: Two-level description of Turkish morphology (K.Oflazer). --p.155-166: Recent research in France: a report and bibliography (M.Juillard).

Terminology, *vol.1 no.1 (1994)*. Contents: p.7-16: Terminology: custodian of knowledge and means of knowledge (Juan C.Sager). --p.17-40: Quelques caractéristiques du vocabulaire de l'acupuncture (J.C.Boulanger and Gaétane Lavigne). -- p.41-60: Data elements in terminological entries: an empirical study (Sue Ellen Wright and Gerhard Budin). -- p.61-96: Positional and combinational characteristics of terms: consequences of corpus-based terminography (Blaise Nkwenti-Azeh). -- p.97-102: Towards a common vocabulary for classification and definition (Pieter F.de Vries Robbé and Frank J.Flier). -- p.103-120: Linguistic representations of concepts in Japanese and English complex noun terms (Kyo Kageura). -- p.121-136: Management of terminology in an MT environment (Marie-Claude L'Homme). -- p.137-146: Terminology for quantities and units in International Standards (Anders Thor). -- p.147-170: Automatic recognition of complex terms: the TERMINO solution (Andy Lauriston). -- p.171-180: Résumé d'une lecture terminologique de *Translation and Meaning Part 2* (Paul Wijnands). -- p.181-192: Applied terminology: a state-of-the-art report (Lynne Bowker). -- p.195-201: Terminology in Canada (Malcolm Williams). -- p.202-204: Pragmatists united: the Deutscher Terminologie-Tag e.V. (John D. Graham). -- p.205-208: Associazione Italiana per la Terminolgia (Giovanni Adamo and Laura Bocci).

Books

Sager, Juan C.: **Language engineering and translation. Consequences of automation.** Amsterdam/Philadelphia: John Benjamins Publ.Co., 1994. xx, 345 pp. (Benjamins Translation Library, vol.1) ISBN: 90-272-2139-1.

Schütz, Jörg: **Terminological knowledge in multilingual language processing**. (Studies in Machine Translation and Natural Language Processing, vol.5) Luxembourg: European Commission, 1994. vi,120 pp. ISSN 1017-6568. [For further details see elsewhere in this issue]

Language industries atlas. Edited by P.M.Hearn and D.F.Button. Amsterdam, Oxford, Washington, Tokyo: IOS Press, 1994. xviii,406 pp. ISBN: 90-5199-148-7. £55.00

Conference proceedings

Globalization Symposium, San Jose, CA 15-17 March 1994. Chêne-Bougeries: Localisation Industry Standards Association, [1994].

32nd Annual Meeting of the Association for Computational Linguistics. Proceedings of the Conference, 27-30 June 1994, New Mexico State University, Las Cruces, New Mexico, USA. [ACL: 1994.] xiii,339pp.

Language Engineering Convention, CNIT, La Défense, Paris, July 6-7 1994. Abstracts, compiled by Leeann Jackson-Eve. [Edinburgh: ELSNET, 1994.] 120pp.

Proceedings of ROCLING VII. R.O.C. Computational Linguistics Conference VII (1994). National Tsing-Hua University (R.O.C.): 1994. viii,292pp.

SNLR: Proceedings of the International Workshop on Sharable Natural Language Resources, 10-11 August 1994, Nara Institute of Science and Technology, Ikoma, Nara, Japan. Edited by Yuji Matsumoto and Takenobu Tokunaga. [Nara: 1994.] 183pp.

Reports

Torrens, Antoni G.: **MT evaluation and quality benchmarks**. A discussion paper (Final version). Luxembourg: Commission of the European Communities, May 1993 (Reprinted December 1993) v,89 pp. [For further details see elsewhere in this issue.]

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.