MT News International

Newsletter of the International Association for Machine Translation

ISSN 0965-5476 Issue no.11, May 1995

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

Published in the United States by Jane Zorrilla

SYSTEMS and PRODUCTS

Systran for Windows Is Here!

[Press release, 27 April 1995]

La Jolla. – Systran Software, Inc., standard-bearer of the most time-honored name in natural language technology, has just unveiled a suite of full-scale machine translation products for the PC. Systran Professional for Windows brings all the power of Systran's patented mainframe technology to the desktop, offering the quality long-term solutions that have made Systran the industry leader for nearly three decades. Systran president Dennis Gachot takes pride in announcing that "this entire world-renowned system is now available to almost anyone."

Five of Systran's 27 language combinations are now ready to ship: English into Spanish, French, German, Italian, and Portuguese. Six more will be released in the course of 1995: Spanish, French, German, Italian, Portuguese, and Japanese into English. Others in preparation include Korean, Chinese, and Russian.

Systran Pro offers all the same capabilities as its mainframe predecessor--and a lot more. The massive, highly refined dictionaries, long recognized as the biggest and the best in the machine translation business, ensure in-depth coverage in a broad range of subject areas. In addition, the user-friendly environment lets customers add their own terms and expressions with ease. Dictionary maintenance is streamlined because it is a true multitarget system. For this same reason, the software is ideal for translating the documentation that enables companies to penetrate many markets simultaneously around the world—just as Systran's veteran customers have been doing for years.

Like its forerunner, Systran Pro preserves all the format of the original document, which can be prepared using any of the leading word processors or as SGML. By capturing the page layout, graphics, tables, and fonts, this software can save as much as half the cost of publishing a translation. When it is integrated with tools that identify previous translations from a stored database, it can help to cut costs even more.

Although Systran Pro inherited its know-how from a long tradition, it is right at the technological edge: a native 32-bit Windows application, it runs in 8 Mg RAM (16 Mg recommended), and it is ready for Windows 95 and Windows NT. There is also a network version that can be used as a server in large operations. The standalone version runs under Windows on a PC-486 and sells for $1,495 per language combination.

Systran Software also offers on-line access to its products via modem. In addition, the company provides complete localization services.

For further information, contact: Stephen Dakis (Tel: 619/459-6700; Fax: 619/459-8487; Email: info@systranmt.com)

Intergraph announces TRANSCEND

[Press release]

Huntsville, Ala. (January 1995) -- Intergraph $^{(R)}$ Software Solutions is breaking the language barrier in today's fast-paced business environment with the introduction of Transcend^{TM}, a 32-bit, natural language translator for Windows^{TM}-based PCs. The new software product translates text to and from common European languages while preserving the format of the original document. For ease-of-use, it is designed to either act as a standalone translator or to be accessed directly from a pull-down menu in a word processor. Unlike lightweight PC translators on the market today, Transcend uses robust linguistic technology that was previously only available on high-end mini- and mainframe hardware running software that cost thousands of dollars.

Improves Global Communication

"The globalized economy presents businesses with a greater communication challenge than ever before," says Intergraph's Deane K. Dayton, Ph.D., Executive Manager, Natural Language Products. "Adding Transcend to your desktop provides quick translations with professional results. It's ideal for small- and medium-sized businesses that want to compete effectively in international markets. The obvious benefits are the time and cost savings in translating documents, manuals, e-mail and other daily business correspondence".

Professional Quality Translations

Transcend has been derived from Intergraph's full-blown UNIX$^{(R)}$-based machine translation system called DP/Translator. The DP/Translator translation engine has been seamlessly incorporated onto the Windows desktop and makes use of extensive dictionaries and a sophisticated set of linguistic rules. The software has been designed for ease-of-use and operates standalone or can be integrated with common word processors such as Microsoft$^{(R)}$ Word or WordPerfect$^{(R)}$.

Transcend gives the same translation quality as used by companies such as McDonald's and AT&T Business Translations, and on CompuServe's MacCIM Help Forum.

Product Features and Benefits

"We're proud that Transcend has powerful features and a low cost that enables individuals and businesses to communicate straight from their desktop," says Dayton. "The ability to quickly translate a document, polish it with minor post-edits, and then distribute it opens many new opportunities for enhanced global communication."

Transcend provides tools that:

* utilize Intergraph's new word processor integration feature
* are Microsoft Office Compatible
* preserve the layout and type specifications of the original document
* integrate inside Word and WordPerfect
* are compatible with several applications and file formats (QuarkXPress, FrameMaker, AmiPro, SGML, ASCII and others)
* search the document, identify unrecognized words and phrases, and perform full-sentence translations

Product Availability

Initial release of Transcend (First Quarter, 1995) includes the language directions: English-Spanish, Spanish-English, English-French, and French-English. Later releases will include: English-German, German-English, English-Italian, and English-Portuguese. The introductory price will be $295 with a list price of $495.

Hardware requirements:

IBM/compatible PC, 386 or higher
Windows 3.1 or later or Windows NT
8 MB RAM, and 15 MB of memory
For integration with Microsoft Word 6.0 or later
For integration with WordPerfect 6.0 or later

For further information on Transcend telephone 1-800-222-9242 (within the U.S. and Canada), or 1-205-730-9832 (elsewhere); Email inquiries to: transcend@ingr.com.

Intergraph Translation Product Nominated for Major Award

[Press release]

Huntsville, Ala. (Apr. 13, 1995) -- Intergraph® Software Solutions announced today its nomination to the 1995 Computerworld Smithsonian Awards in the science category.

CompuServe of Columbus, Ohio nominated Intergraph for its machine translation technology. The technology developed at Intergraph provides rapid translations from one human language to another, enabling CompuServe to offer multilingual communication on computer bulletin boards and forums.

"The machine translation industry is still young. Being nominated for such a prestigious award means that this technology is already making a difference in people's lives and that our customers feel we represent the best in the industry," says Rich Buchheim, Intergraph

Executive Vice-President for Information Management and Solution Engineering. "We are extremely pleased with the response that our PC-based application, Transcend™ Natural Language Translator, has received. It has made professional-quality machine translation more accessible and added awareness to how machine translation offers expanded global communications."

Widely recognized as the premier awards program in the information technology industry, the Computerworld Smithsonian awards honor corporations, groups and individuals who are using technology to create positive change in society.

Intergraph's machine translation technology is incorporated into Transcend, Natural Language Translator. Transcend translates text to and from common European languages while preserving the format of the original document. The software can be integrated with common word processors such as Microsoft® Word or WordPerfect®.

Background Information

A member of the Fortune 500, Intergraph Corporation (Huntsville, Ala.) is the world's largest company dedicated to supplying interactive computer graphics systems. Products range from point solutions, meeting individual and departmental needs, to integrated, enterprise-wide systems.

Noted for delivering interoperable systems and applications, Intergraph bases its products on Windows, Windows NT, and UNIX operating systems.

As one of the company's business units, Intergraph Software Solutions (ISS) develops and markets integrated software for the technical desktop the combination of compatible technical applications and personal productivity tools in a single desktop computer. Technical applications include computer-aided design, engineering, analysis, manufacturing, publishing and earth sciences. ISS also provides core system software, high-end applications, and training, consulting, and implementation services.

Established in 1979, the CompuServe Information Service provides its worldwide membership of 2.2 million with databases and services to meet both business and personal interests. CompuServe can be accessed by any modem-equipped personal computer using general communications software. In addition to the CompuServe Information Service, CompuServe Incorporated provides frame relay, wide area networking services, electronic mail, entertainment and business information services to consumers and major corporations worldwide. CompuServe is an H&R Block (NYSE:HRB) company.

Intergraph is a registered trademark and Transcend is a trademark of Intergraph Corporation. Other brands and product names are trademarks of their respective owners.

For further information contact: Susan Moore (Tel: +1 205 730 3315; Email: sjmoore@ingr.com). Intergraph Corporation is on-line on the Internet at: http://www.intergraph.com for additional product information.

Logos targets freelance translators

[Press release - from LISA Forum 4:1]

LogosClient dial-up software to let freelancers access regional MT servers via PC's.

Jens Thomas Lück, CEO of Logos Corporation, has announced a major new strategy aimed at the freelance translator. At the heart of this strategy is a new PC-based LogosClient with dial-up feature that permits users to access a remote server for machine translation while doing pre-editing, lexical maintenance, and post-editing off-line.

LogosClient is currently integrating in MS Word, WordPerfect and AmiPro. Users can import or create documents in these standard word processors, pre-edit the text to be translated (if desired), do off-line dictionary work, and then, without leaving these editing environments, instruct LogosClient to upload the document and associated lexical updates to the server for processing.

LogosClient automatically dials up the server, uploads the data, gets an estimate of job completion time, and then disconnects. Upon job completion, LogosClient retrieves translated output for off-line post-editing. Formatting of the source document is faithfully reproduced in the translated version. Connect time and line charges for the entire procedure are expected to be negligible.

Logos Servers will be placed at various locations around the country, in most cases at a regional translation bureau. Preparations are being made for certain translation bureaus to function as customer support centers as well.

LogosClient dial-up has successfully completed alpha testing and is about to enter beta testing in an East cost/West coast hookup. The product should be available for general release by the second quarter of this year. Logos expects to make the LogosClient software available to freelancers at next to no cost. Charges to translators will be on a per usage basis.

For further information contact: *USA:* Logos Corporation (Attention: Christophe Mosing), 200 Valley Road, Suite 400, Mt.Arlington, New Jersey, USA. (Tel: +1-201-398-8710; Fax: +1-201-398-6102; Email: cmosing@logos-usa.com); *Europe:* Logos GmbH (Attention: Friederike Bruckert), Mergenthalerallee 79-81, D-65760 Eschborn/Ts., Germany (Tel: +49-6196-59030; Fax: +49-6196-590215; Email: bruckert@logos-usa.com)

The Logos Product Line

[Publicity sheet, February 1995]

Logos Server for SunR SPARCstations™ (Model 5/20)
with the following language pairs:
English → Spanish
English → French
English → German
English → Italian
German → English
German → French
German → Italian

Logos is integrated with the leading word processing and desktop publishing software packages.

Depending on customer requirements and configuration, LogosClient and partner products are easily integrated in a network and may thus support access to Logos Server directly.

LogosClient for Windows is a Microsoft Windows based application program to support access to the Logos™ Translation Server Software in a corporate local area network environment. This access can take place from the menus of the leading word processor software packages in Microsoft Windows.

LogosClient for OSF/Motif. This client product is an OSF™/Motif® based application program to support access to the Logos™ Translation Server Software in a corporate local area network environment. The access can take place from the menus of the leading UNIX desktop publishing software packages.

Partner Products

EUROLANG *Optimizer*™ for Logos. The combination product of *Optimizer*'s Translation

Memory with the Logos Intelligent Translation System is fully integrated with Word for Windows 6.0 (integration with FrameMaker for UNIX will be released shortly) enabling users to work in their familiar software environment. The *Optimizer* software runs on Sun® SPARCstations™ or Pentium PC (server) and *Optimizer* PC clients in a network.

XL8® for Logos (GlobalWare®). The combination of XL8® PC based Translation Memory and process management tools with all functions of the Logos Server Software, enables the successful management of localization products. File processing may range from preformatted technical documentation to online help systems and text in graphical user interfaces.

Features of Logos Server:

Structure and size of Logos dictionaries (unlimited expansion):

250 predefined virtual Logos Subject matter Dictionaries

250 virtual Subject Matter Dictionaries freely definable by the user

Number of basic entries:

English source language: currently more than 50,000 basic entries (stems) incl. more than 3,000 in virtual Subject Matter Dictionaries

German source language: currently more than 75,000 basic entries (stems) incl. more than 15,000 in virtual Subject Matter Dictionaries

Structure of dictionary entries:

currently a maximum number of characters per source term of 34 characters for ALEX entries

maximum number of words in the source term: 10 or limited by maximum number of characters

currently a maximum number of characters per target term of 48 characters for ALEX entries

maximum number of words in the target term: 10 or limited by maximum number of characters

Dictionary selection:

a maximum of five Company Dictionaries can be selected hierarchically

a maximum of five Subject Matter Dictionaries can be selected hierarchically

three-letter code representing the Company Dictionary

three-digit code representing the virtual Subject Matter Dictionary

Structure and size of the semantic database (unlimited expansion)

three-letter code representing the Company Dictionary

a maximum of five Company Dictionaries can be selected hierarchically

currently SEMANTHA offers seven rule templates for German source language and eight rule templates for English source language.

Number of semantic rules:

English source language: currently more than 11,000

German source language: currently more than 20,000

Input documents:

Maximum length of path names: 100 characters

Speed of translation

The speed of translation is 1 or 2 pages (letter format) per minute. This speed is dependent upon the system configuration, amount of formatting information and the length and complexity of the sentences in the text.

System configuration

Hardware:

System: Sun SPARCstation

Processor: SPARC™, Model 5/20 or comparable, or higher performance
Main memory: minimum of 32 MB RAM
Free disk storage:
160 MB for English source language
210 MB for German source language
300 MB for both source languages
depending on the translation volume: further 200 MB necessary
Network protocol: LAN TCP/IP (optional)
Network interface: THIN/THICK ETHERNET (optional)
Magnetic tape drive: 150 MB QIC or 8mm DC
High-resolution monitor

Software:
Operating system. Server:
Sun OS 4.1.3 with System V extension or Solaris® 2.3
Open Windows™ 3.0 or OSF/Motif

Further Logos Application Software:
Client: LogosClient for Windows™
LogosClient for OSF/Motif (UNIX)
Partner products: Optimizer™ for Logos (Eurolang®)
XL8® for Logos (GlobalWare®)

The following formats are currently supported:

Interleaf® 5.3 (Forced ASCII 5.3); FrameMaker® 3.1 (MIF 3.1); Word for Windows™ 2.0b, 6.0 (RTF 1.2); AmiPro® for Windows™ 3.01 (RTF 1.2); WordPerfect® 5.2 (DOC (WordPerfect) 5.2); IBM Editor (SCIPT A, SCRIPT); DCA (RTF); nroff/troff (UNIX); XL8 1.6 (LTX); SGML (provisional); Logos API (LTX).

Globalink Available on Minitel

[Press release]

Fairfax, VA. (January 4, 1995) -- Globalink, Inc. (ASE: GNK) announced it has signed an agreement to license its language translation software technology to Meta International and MEM Tilicom & Riseaux of France. The licensing agreement enables the two systems integrators to offer multilingual access to the more than 6,000,000 users to Minitel, the French on-line service.

Under the Meta and MEM multi-year agreements, which have a combined estimated value of $820,000, each company will license the Globalink translation engine to be integrated with their French user interface to Minitel. This allows non-French speaking users on-line access to Minitel\022s databases and services. Initially, this interface will be available in English with the anticipation of adding more languages as the need arises.

Meta International, a French telecommunications systems integrator, has offices in Europe, the U.S. and Latin America. Meta\022s product, Traductel, provides PC users Minitel emulation plus on-line translation in English to and from French when making a system inquiry.

MEM Tilicom & Riseaux, a French research and development company in telecommunications, is a UNIX systems integrator. MEM will be providing a solution on-line, with a UNIX server, for automatic translations with specialized dictionaries for French industry as well as access to UNIX users on Minitel.

Globalink's President, Michael E. Tacelosky, stated, "Globalink is committed to overcoming language barriers and encouraging people to communicate with each other around the world. Certainly, giving access to 6,000,000 on-line users in France and also a few million foreign telecommunications users worldwide adds to this goal."

Both Meta and MEM said that using Globalink's technology not only gives non-French speaking users access to Minitel but also provides a unique opportunity to broaden exposure to the multilingual global community well beyond the French-speaking world.

Globalink, the worldwide leader in language translation software, offers a wide range of translation tools to meet the specific price and functionality requirements of the consumer, business and professional markets. Versions of Globalink's software products – Language Assistant™, Power Translator ®and Power Translator Professional – are available for Spanish, French, German or Italian on a wide range of operating platforms including Windows, DOS and Macintosh (R).

For information about Globalink products or other language translation services available from Globalink (9302 Lee Highway Fairfax, VA 22031-1208 USA), call 800-255-5660 or 703-273-5600; Fax 703.273.3866

ATLAS machine translation service offered on PCCs

[Extracts from AAMT Journal no.7, June 1994.]

Since our company [Fujitsu] pioneered a machine translation service on the personal computer communication [PCC] network NIFTY-Serve in early October 1990, many other similar MT services have been offered, and have been joined by post-editing services. A commercial MT system is applied for the machine translation service on PCC so that the translation quality is equal to that of any general machine translation use.

The full process is automated, from the collection of the translation requests to the transmission of the translation result. Through the electronic mail system of the PCC, the documents requested to be translated are received, and the translation result is transmitted.

New version of HICATS from Hitachi

[From AAMT Journal no.8, September 1994]

The Hitachi machine translation (MT) system "HICATS" is an integrated translation support system with bi-directional (Japanese↔English) translation functions.

Hitachi has been conducting MT research since the latter half of the 1970's, and it started marketing the "HICATS" MT system in 1986. At first, systems on mainframe computers were marketed, and from 1989 on, workstation systems were developed. Thereafter, high quality, user-friendly MT systems were developed by supplementing the various functions, and by improving and expanding the grammatical data and dictionary data.

Here, we introduce an MT system that runs on the new high-performance workstation "Hitachi Creative Station 3050-RX group".

* **Easy operation**

Corresponds to X Window in accordance with the industry standard OSF/Motif. The MT output can be edited and proof-read on a multi-window screen. A pop-up menu and a mouse-driven man-machine interface simplify such functions as word selection and correction, command selection, translation range specification, and screen scrolling, etc.

* **Easy term-dictionary compiling function**

For words that require user-specified translation equivalents, a user dictionary is provided for registering English equivalents chosen by the user. An easy to operate "Dictionary maintenance function" is provided to support user dictionary registration.

*** Learning function for facilitating user dictionary registration**

In the case of Japanese-English translation, English equivalents post-edited by the user are sorted out and registered automatically in the user dictionary. Moreover, a post-edited verb and its nominal complements, etc., are registered in the user dictionary as a single entry. In the case of English-Japanese processing, English equivalents of a specific word are selected and registered in the user dictionary. User dictionary registration becomes much simpler when the learning function is used.

*** Highly developed Japanese sentence analysis function**

Consider the following Japanese sentence:

1. HICATS/JE ni yotte honyaku shita ronbun to manuaru wo henshuu shita
   BY TRANSLATED THESIS AND MANUAL OBJ EDITED

The ambiguity here is: does the modifier [*HICATS/JE ni yotte honyaku shita*] modify the NP [*ronbun to manuaru*] (2, below) or only the NP [*ronbun*] (3)?

2. [HICATS/JE ni yotte honyaku shita [ronbun to manuaru]] wo henshuu shita
3. [[HICATS/JE ni yotte honyaku shita ronbun] to manuaru] wo henshuu shita

Furthermore, there is the question of whether the adjunct [*HICATS/JE ni yotte*] is an adjunct to the object NP [*ronbun to manuaru*] (5) or to the matrix verb [*henshuu suru*] (6).

5. [HICATS/JE ni yotte honyaku shita ronbun to manuaru] wo henshuu shita
6. HICATS/JE ni yotte [honyaku shita ronbun to manuaru] wo henshuu shita

The Japanese sentence analysis function checks the ambiguity of such Japanese sentences. 15 kinds of such diagnostic items for complex sentences and ambiguous words, etc., are provided.

*** The translation output option selects the translation sentence pattern**

For instance, a subject-less Japanese sentence is usually translated into a passive English equivalent. However, in specific kinds of documents, such sentences are better translated into their imperative equivalents. In other words, supplementing the subject "you" may be more appropriate during the translation of the original Japanese sentence. The user can specify the translation method at the outset by means of the translation output option, and this reduces the post-editing effort. 23 kinds of translation output options are provided.

*** Translation accuracy is improved by utilizing co-occurrence relationships**

The source language sentence is analyzed accurately by using knowledge regarding the simultaneous occurrence (co-occurrence relations) of given words. Simultaneously, the appropriate equivalent of each word is selected and an easy-to-understand translation is generated.

Japanese-English translation sample:

(Translation of "*ageru*")

sono kuruma wa supiido wo ageta
THAT CAR TOPIC SPEED OBJECT INCREASED
→ The car increased speed
kare wa te wo ageta
HE TOPIC HAND OBJECT RAISED
→ He raised a hand

English-Japanese translation sample:

(Deciding the meaning of "*bring*")

He brought me the flowers on the table
→ kare wa watakushi ni teeburu no ue no hana wo motte kita
HE TOP I TO TABLE OF ABOVE OF FLOWERS OBJ BRING-CAME

He will bring his daughter to the party tomorrow

→ kare wa ashita kare no musume wo paatii ni tsurete daroo

HE TOP TOMORROW HE OF DAUGHTER OBJ PARTY TO ACCOMPANY WILL

*** Compatibility with various file formats**

Documents prepared with OA software "OFIS-EX series" operating on Hitachi workstations 3050 and 3050RX can be directly used as source language translation texts. Documents containing figures as well can be translated as well without any pre-editing to mask those figures. Translations can be prepared with figures just as in the original text. Moreover, the translation of files input in te UNIX format or the MS-DOS format can be output as UNIX and MS-DOS files. These functions reduce input time and ease the editing work.

*** Electronic dictionary**

An electronic dictionary which contains both the Kenkyusha's new Japanese-English dictionary and the new English-Japanese dictionary is supplied. The electronic dictionary is useful during post-editing since the dictionary can be referred to, and dictionary entries can be copied into a bilingual editor. Furthermore, an English spell-checking function is included. Accordingly, the misspelling of the English input can be checked easily and corrected.

*** Domain specific terminology dictionary**

Domain-specific terminology dictionaries in eight fields are provided:

(1) Information processing
(2) Electrical and electronics engineering, and communication
(3) Engineering works and architecture
(4) Automobiles, railways, ships, and aviation
(5) Natural sciences
(6) Biology
(7) Mechanical engineering
(8) Chemical industry

Performance and prices

Various items of hardware and software are offered. It is possible to choose according to the user's operation mode. Some examples:

-With 3050RX-200:
Price starts from JY 3,500,000
Translation speed =
Japanese-English translation:
About 44,000 words/hour
73 pages/hour (600 words/page)
English-Japanese translation:
About 57,000 words/hour
228 pages/hour (250 words/page)

-With 3050RX-330:
Price starts from JY 7,700,000
Translation speed =
Japanese-English translation:
About 110,000 words/hour
183 pages/hour (600 words/page)
English-Japanese translation:
About 140,000 words/hour
560 pages/hour (250 words/page)

Inquiries to: Product Planning Headquarters, Computer Business Headquarters, Hitachi Co., Ltd. (Tel: +81-3-3763-2411, extension: 3280)

HONYAKU KOBO from Panasonic

[From AAMT Journal no.8, September 1994]

Panasonic English to Japanese automatic translation system Honyaku Kobo is born! The Panasonic English to Japanese Automatic Translating System "Honyaku Kobo" does not simply do translation processing; it is a total documentation system which reads in English documents and completely processes the translated document.

The main features of "Honyaku Kobo"

*** Layout Function**

"Honyaku Kobo" can complete and print translated documents with exactly the same layout as the original English documents, including any figures, formulas, or graphs. This "Layout" function is one of the outstanding features of this system. Using this function makes cutting and pasting graphics, etc., onto the translated text unnecessary. In this way, "Honyaku Kobo" makes it possible to greatly reduce the total working time required to prepare the translation.

*** High speed and High quality translation**

"Honyaku Kobo" can translate 16 thousand words per hour (using Sparc Station 10). Furthermore, high quality translations are obtained by adopting the "Tree Structure Conversion Method" which has the most grammatical descriptive power. Documents can be efficiently translated by translating important sections thoroughly and rapidly translating other portions that need to be translated quickly.

*** Easy operation**

"Honyaku Kobo" considers the user's convenience first. All operations can be carried out by using a mouse. A GUI which uses Japanese Motif provides a user-friendly display. Moreover, there are many tools for the modification of the translated documents.

System configuration:

Hardware: main unit - Sparc workstation; scanner - IS50 (ADF attached) from Ricoh; printer - LBP Panasonic KX-P4630.

Software: user interface - Motif; OS - Unix; GUI - X Window; Translation software/dictionaries; Character recognition software.

Dictionaries: basic - 70,000 words; technical - 3,000 to 40,000 words (Information technology, Electrical engineering, Mechanical engineering, System control engineering, Heat engineering, Energy engineering, Transportation & communication engineering)

For further information contact: Nagase & Co., Ltd. Translation bureau (Tel: +81-3-3665-3060 or 3396), Kyushu Matsushita Electric Co., Ltd. Takahashi (Tel: +81-92-477-1548)

Transwise: Machine Translations from Finnish to English

[Press release 7th March 1995]

Two Finnish companies, Kielikone Ltd and Trantex Ltd, established a joint venture company called Transwise on January 1, 1995. The new company offers machine translation services from Finnish to English. The translations are produced with the Translator's Workstation, developed by Kielikone. Machine translation services are targeted especially at customers who need basic translations from Finnish to English quickly and at low cost.

Transwise is a subsidiary of Kielikone and operates at the same location as Kielikone in Helsinki. Kielikone holds a 75% share in Transwise and is responsible for producing the translations. Trantex markets the services and participates in system development.

The area of machine translations is developing rapidly. As markets are becoming increasingly global, more organizations and individuals need a fast, attractively priced means of having their texts translated into other languages. The Translator's Workstation, developed by Kielikone, is the first functional machine translation system for Finnish. Kielikone also licences the Workstation, so that customers can either buy translation services or purchase the system license.

Kielikone specializes in developing language software. In addition to the machine translation system, its product range includes electronic dictionaries and grammar and style checker software. Kielikone has also developed a high-quality morphological analyzer and a parser for Finnish.

Kielikone, which was sponsored originally by Sitra and later by Technology Development Centre, has been developing its machine translation system since 1987. In addition to Trantex, Nokia Telecommunications and Rautaruukki have participated in the development work. The machine translation system uses a general dictionary of over 50,000 words, plus special dictionaries that can be created and compiled for each customer or project. The system can translate several sentences per second.

Trantex is one of the largest localization agencies in Europe. The company was established in 1983, and by January 1995 its staff had grown to 82 people and its net sales to FIM 19 million (USD 4 million). Trantex specializes in technical writing and in localizing software, help systems and manuals. Through its subsidiary, Bitwit, it also offers training and product support services. Trantex is a Microsoft Solution Provider and an authorized Drake Test Center. A network of international partner companies ensures a large language capability.

For more information please contact:

Kielikone Oy, P.O. Box 126, 00211 Helsinki, FINLAND (Tel +358 0 682 02 11; Fax +358 0 682 01 67; Harri Arnola, Email: harri@kielikone.fi; Petteri Suoranta, Email: petteri@kielikone.fi)

Trantex Oy, Ahventie 4 B, 02170 Espoo, FINLAND (Tel. +358 0 613 35 00; Fax +358 0 613 35 390; Harri Pohja, Email: harrip@trantex.fi)

New versions of Eurolang Optimizer

[Press release in LISA Forum 4:1]

EUROLANG Optimizer™ 2.1 will be released in March 1995. The main new functionalities of this new release are:

Efficiency of the pre-translation server improved
Term and sentence databases
Man Machine Interfaces enhanced
Smart display conflict function for terms and sentences databases
Possibility to activate pre-translation from translator's PC running Windows 3.11
Undo, Revise Paragraph on the translator's workstation

EUROLANG Optimizer™ Workstation 2.1 will be released in the 2nd quarter of 1995. It will run on Windows NT 3.5 with Word 6.0.

EUROLANG Optimizer™ for Logos™ A combined product: machine translation and translation memories. The merging of machine translation and translation memory technologies brings a powerful new productivity tool for translators. The synergy of two complementary technologies can yield a result that is unarguably greater than the sum of the parts. This appears to be the case with the deep integration currently underway between EUROLANG Optimizer and the Logos Machine Translation system. Rarely have two technologies complemented each other so successfully. What appeals most is the friendliness of the product, particularly the control it gives the user regarding what source of electronic assistance he/she is offered at any given time in the translation process.

Optimizer for Logos is designed with a range of options all aimed at helping the translator. These include translation memory, on-line document glossaries, and machine translation, all of which the translator can make use of, or not, on a sentence-by-sentence basis as appropriate.

If you are already familiar with the EUROLANG Optimizer, this combination of these products will appear obvious. When a sentence has a Perfect Match Memory Translation, it is proposed. For every other sentence Optimizer for Logos proposes a Machine Translation, and possibly a Fuzzy Match Memory Translation and Technical Term assistance. The translator can validate the proposals, or not, and is free to modify them.

The potential gains in productivity with Optimizer for Logos is quite impressive given the track record of these two technologies taken separately.

KIT-FAST system available

Wilhelm Weisweber

[From WWW files at Technical University Berlin]

The project KIT-FAST was (from June 1985 to December 1992) a basic research project in MT within the project group KIT. It was the Berlin component of the complementary research to Eurotra-D, which was itself the German part of the European Community MT project Eurotra, and it was funded by the Federal Ministry for Research and Technology.

The major task of the complementary research was to check recent linguistic theories and AI methods and make them available for the specific problems of MT. The first phase of the Berlin project (KIT-NASEV) dealt with syntactic issues. NASEV is an abbreviation for the project title *New Algorithms for Analysis and Synthesis* in MT. A constructive version of GPSG (Generalized Phrase Structure Grammar) was developed and implemented as the syntactic component of an MT system.

The second phase (KIT-FAST I) concentrated on sentence semantic problems of translation. FAST is an abbreviation for *Functor-Argument-Structure for Translation*. The project title was *Transfer and generation on a sentence-semantic level*. The project designed and implemented a representation formalism for sentence-semantic information and corresponding components for semantic analysis, transfer and generation on the basis of term-rewriting, which have been connected with the syntactic component of the MT system.

The third phase (KIT-FAST II) was concerned with the problem of anaphoric interpretation in MT. The title of the project was *Anaphora resolution in MT*. The project developed a method for anaphora interpretation that takes a whole variety of different factors into account. The factors concern the structural prominence of an antecedent candidate as well as the conceptual consistency of a text. The latter is determined with the help of predefined background knowledge and a representation of the text content, which are represented in the TBox (terminological knowledge) and ABox (assertional knowledge), respectively, of a KL-ONE based knowledge representation (KR) system. The KR system has been developed independently by the neighbour project KIT-BACK.

The KIT-FAST experimental MT system is transfer-based and translates written

German texts into English sentence by sentence. The translation of a sentence consists of morphological, syntactical, semantical and conceptual analysis, transfer, generation and morphological synthesis. The algorithms for morphological analysis and synthesis are based on the SUTRA system (a module of the HAM-ANS hotel information system). The syntactic analysis is realized by a GPSG parser, which interprets ID rules, LP statements and metarules directly. The semantic and conceptual analysis, the transfer as well as the generation is realized by one algorithm on the basis of term-rewriting (known from the automatic proof of equations).

After the semantic analysis of a sentence the resulting FAS expression is conceptually analysed, i.e. it is mapped onto an expression of the ABox-Tell-Language (ATL), with the help of which the contents of the sentence is added to the representation of the text content in the ABox of the BACK system.

Our first step towards the translation of German texts instead of single sentences was to interpret anaphoric relations in the source language. For that reason an algorithm for the evaluation of anaphoric relations has been developed and implemented. This algorithm uses the textual and background knowledge in order to determine the structural prominence of an antecedent candidate and its consistency with the anaphor.

The evaluation component of the MT system takes a FAS expression of the source language as its input and looks for antecedents in the same and preceding sentences. After evaluating anaphorical relations the FAS expression is actualized, i.e. the parts of the FAS expressions corresponding to an anaphor and its antecedent are made to refer to the same ABox object. An actualized FAS expression for a source language sentence is transferred into a target language FAS expression, from which the corresponding target language sentence is generated.

The MT system employs two textual representations. One for representing the structural information of a text and another for representing the text content. The textual representations are constructed incrementally from the sentential ones during translation. In principle a textual representation is needed on every level, but this would lead to redundant representations on the syntactic and semantic level. For that reason we decided to take the more general semantic level (FAS) for the representation of structural aspects of the text.

The *components of the MT system*: morphological analyser based on the SUTRA system; GPSG parser for direct interpretation of ID rules, LP statements and metarules; term-rewrite rule interpreter for semantic and conceptual analysis, transfer and generation; morphological synthesizer based on the SUTRA system; module for the evaluation of anaphoric relations; the knowledge representation system BACK; tools for the development of lexicons, grammars and term-rewrite systems.

The *linguistic data*, developed in order to translate a German text (The Proposal of the European Commission for the ESPRIT Programme), comprise: a German grammar (GPSG) with 22 main categories, 34 features, 22 aliases, 76 ID rules, 23 LP statements, 5 metarules, 23 FCRs, 265 lexical entries (stem forms), 134 term-rewrite rules for semantic analysis (German), 37 term-rewrite rules for conceptual analysis (German), 248 term-rewrite rules for transfer (German → English), 182 term-rewrite rules for generation (English), 8 factors for the evaluation of anaphoric relations in German (agreement, binding, proximity, preference for the semantic subject, topic preference, identity of roles, negative preference for free adjuncts, conceptual consistency). The predefined background knowledge comprises selectional restrictions. About 100 sentences were successfully tested with the help of the MT system.

Implementation of the MT system is in Quintus-Prolog 3.1 (commercial software) and SWI-Prolog 1.9.5 (public domain software). Both Prolog dialects are running on Sun workstations under SunOS and AT compatible PCs under DOS (Windows 3.1). The MT system is tested for Quintus- and SWI-Prolog under SunOS and under SWI-Prolog under Windows 3.1 and needs about 10 MB of hard disk space.

In order to get the software for the MT system running on AT compatible PCs under DOS (Windows 3.1) see http://www.cs.tu-berlin.de/~ww/mtdos.html.

If you are interested in receiving the software for the MT system for Sun workstations under SunOS see http://www.cs.tu-berlin.de/~ww/mtsun.html.

The list of available KIT reports can be found at http://www.cs.tu-berlin.de/~kit/reportliste/kitlistehtml.html.

Further Information: Wilhelm Weisweber, Technical University of Berlin, Department of Computer Sciences, Institute for Software and Theoretical Computer Sciences (ISTI), Functional and Logic Programming (FLP), Sekr.: FR 6-10, Franklinstr. 28/29, D-10587 Berlin-Charlottenburg, Federal Republic of Germany (Tel: +49-30-314-73608; Fax: +49-30-314-73622; E-mail: ww@cs.tu-berlin.de; WWW: http://www.cs.tu-berlin.de/~ww/)

Compiling Dictionaries for MT Systems: a Technology for PARS

Michael Blekhman

Lingvistica '93 Co. and Laboratory for Machine Translation, Kharkov Polytechnical University, have developed what we call The Automated Dictionary Creation Technology. At present, our aim is to create dictionaries for the PARS system, however, in the nearest future, other language pairs will be covered, among which are German-Russian, Russian-Ukrainian, and English-Ukrainian.

According to the technology, the following sources are used to compile the dictionaries:

- existing printed bilingual dictionaries,
- existing electronic bilingual dictionaries,
- real texts.

The following procedure is used to enter words into the target dictionaries from the printed and electronic ones:

(1) scanning or manual entering, if the printed copy quality is poor;

(2) converting into text format; "extra" fragments are deleted automatically and manually, if necessary, such as comments, transcriptions, etc.;

(3) converting into PARS communicative format followed by importing to PARS;

(4) the dictionary obtained is encoded in a batch mode: a special program is applied to attribute grammatical information to each word, according to the "equality principle", i.e. the words are compared with the previously encoded ones; besides, phrases are analyzed, and a word does not acquire any grammatical description if it has been recognized as being an invariable part of the phrase;

(5) system linguist browses the dictionary obtained and applies the automatic encoding facility to the words for which no "prototypes" have been found by the batch processing program; automatic encoding is performed according to the "similarity principle", using the special grammatical index file.

We also compile MT dictionaries on the basis of real texts. The latter are machine translated, and the "new" words, marked as such by the translation program, are entered into the dictionary by the dictionary officer, directly from the screen. This is followed by automatic encoding the word entered.

The above technology has been and is being applied for developing numerous PARS dictionaries, among which are technical, economic, mining, oil/gas, medical, and a large general usage dictionary based on one of the best English-Russian dictionaries by Prof.Mueller.

Here is the list of existing PARS dictionaries and those under way: general, economy, computers, machine building, medicine, microelectronics, geology/mining, patents, ecology, law, oil/gas technology, aerospace engineering.

ANNOUNCEMENTS

General Assemblies of IAMT and EAMT

During MT Summit V (Luxembourg, 11-13 July 1995) there will be general assemblies of the International Association for Machine Translation (IAMT) and of the European Association for Machine Translation (EAMT). The meeting of EAMT is provisionally scheduled to take place at 13.15 on 11th July. The general assembly of IAMT will occur provisionally at 12.30 on the following day, 12th July. Details of arrangements will be given at the conference.

For information about the programme of MT Summit V see the section **Conference Announcements** in this issue.

AMTA to host MT SUMMIT VI

Proposals Invited

In 1997 the MT Summit will again be held in the Americas. Institutions or groups interested in hosting MT Summit VI are invited to contact AMTA President Muriel Vasconcellos (71024.123@compuserve.com) concerning the elements of a proposal.

MTAPI Committee Announces Special Conference

The Machine Translation Application Programming Interface (MTAPI) Special Committee, working with the Standards SIG, will hold a mini-conference at the Hotel Del Coronado in San Diego on Friday, July 7th. The goal of the committee is to create a standard for interaction between independent software vendor (ISV) products and machine translation programs. The Special Committee is made up of both MT vendors and application developers.

An industry-wide interface would provide e-mail, on-line, translation memory and other application users with direct access to compliant translation software. Major MT vendors including Systran, Globalink and Logos have already agreed to support the new standard, and several application developers are looking at integrating MTAPI on a pilot basis.

For more information contact: MTAPI Committee Chairman, Michael Tacelosky, at (703) 273-5600.

PEOPLE ON THE MOVE...

It is with considerable regret that we have to report that **Joseph Pentheroudakis** (Microsoft

Corporation) has found that his growing commitments leave him no time to continue as AMTA regional editor for MT News International. His contributions in the first three years of the newsletter have been immeasurable, decisive and crucial to its success. His input will be greatly missed.

On the design and production side, Joseph has been succeeded by **Jane Morgan Zorrilla**, who is responsible for the new look of MT News International since the last issue. Her email addresses is: 70671.1560@compuserve.com

On the editorial side, we welcome **David Clements** as our new AMTA regional editor. He can be contacted at: Dr. David Clements, Globalink, Inc., 4375 Jutland Drive, Suite 110, San Diego, CA 92117. His fax number is: +1 (619) 490-3684; and his email addresses are: CompuServe: 71530,3476; Internet: clements@globalink.com (or: 71530.3476@compuserve.com)

David Clements is a Senior Linguist at Globalink, Inc. in San Diego, where he has worked since receiving his Ph.D. from the University of California.

DATABASES and SERVICES

ALEP -- Advanced Language Engineering Platform

ALEP Initiative

The Advanced Language Engineering Platform (ALEP) project is an initiative of the European Commission (EC) to provide the natural language research and engineering community in Europe with a versatile and flexible general purpose research and development environment. The EC has put in motion the development of such a language engineering platform. This platform comes with a set of integrated tools and a mainstream and reasonably powerful linguistic formalism. Since the system is open, modular and fully supported, users cannot only use the platform `as is', but also further extend and enhance it and build in their preferred tools.

Natural language processing (NLP) and related projects currently lack a solid, commonly accepted and widely available platform for the development of large scale, professionally designed linguistic resources and applications. As a consequence, researchers and system designers are forced to build the tools and development aids they need from scratch, before undertaking the implementation of what matters most to them, linguistic resources or applications. This situation constitutes a major bottleneck for any serious attempt to build a strong and effective European NLP industry.

The EC has therefore decided, within the Linguistic Research and Engineering (LRE) programme, to invest in a generic formal and computational environment, which can be put at the disposal of Community and national R&D projects in relevant areas, thereby avoiding duplication of effort across research projects. In making widely available the ALEP system, the EC aims to promote cooperation between different research centers and to progress towards portability and re-use of research results.

A typical user of the ALEP system will be either a skilled researcher in computational linguistics or a team of researchers and application designers, who will be provided with a software environment enabling them to produce linguistic descriptions of different languages, for a number of NLP application domains.

Stages of the ALEP Initiative

The development and distribution of ALEP was planned in a number of stages:

1. Preparation and design (1991 - 1991)
2. Development (1992 - 1994)

3. Phase-in (1994 - 1995)

Preparation and Design

In preparation of the initiative to build this general purpose language engineering environment, the EC commissioned a number of feasibility and design studies to prominent European software companies and research centers. The design studies covered topics such as the linguistic formalism, the software environment and the text-handling system.

They addressed the following key issues:

1. An expressive and efficient formalism – the ALEP formalism – and associated interpreter, capable of handling dictionaries, grammars and texts at an acceptable speed; the ALEP system is formalism independent and thus usable for NLP projects with different formal and operational specifications.

2. A user-friendly environment, with a consistent and intuitive graphical interface, for creation and maintenance of grammatical and lexical descriptions, providing a comprehensive set of interactive debugging and testing aids.

3. A Text-handling system, for analysis and mark-up of structured and unstructured texts to be processed by NLP applications.

Great emphasis was placed on the openness and modularity of the system architecture, so that individual components can be developed, replaced and customized by third parties. In addition, one of the key design aims was to ensure reusability of the linguistic resources created with ALEP.

The architecture of the ALEP development system is based on 5 layers:

1. Presentation layer
2. Control layer
3. Task and object layer
4. Application layer
5. Storage layer

The first three layers constitute the ALEP environment itself. This is supplied with a number of tools, applications and basic linguistic resources. New and replacements tools, applications and lingware can be integrated into this environment.

Development Stage

The development stage was split into two cycles. The first cycle produced an initial operational software environment. The main features of this single-user version were:

- graphical user interface;
- overall user environment, including editors, object browsers, etc.;
- linguistic processing tools with appropriate debuggers;
- basic text-handling system;
- preliminary implementation of a lexical database component.

This first operational version of the system is referred to as ALEP-1 and was available by mid-1993. The system proved to be sufficiently stable, and was packaged with user documentation and installation aids, before being made available to selected sites in the second half of 1993, for assessment and testing.

The basic objectives of the second development cycle were:

- enhanced user interface;
- multi-user capabilities and lingware management facilities;
- improved efficiency;
- porting ALEP-1 to another operating system and hardware platform;
- more sophisticated text-handling component.

Phase-in Stage

The final software product has a versatile set of tools, professionally maintained and supported. The EC intends to devote much of its effort during the phase-in stage to the widespread distribution and familiarization of users with the ALEP-2 system. During this stage, training on the system's linguistic and software engineering aspects will be organized. The Commission plans to distribute the system as widely as possible. The system is to be made available at a nominal cost and support will be guaranteed by a professional team of software engineers and computational linguists.

Support and Availability

In the first quarter of 1994 the EC entrusted the maintenance, support and distribution of ALEP to Cray Systems, Luxembourg, under the LRE programme (LRE-62101). These support services are freely available and ensured for all ALEP User Group (AUG) members until the end of 1995.

Supported Actions

It is planned to port ALEP to a range of software and hardware platforms, as well as to integrate results of projects that have used ALEP.

In addition, an ALEP User Group (AUG) has been formed, organizing workshops and training courses for the system.

On-going projects will contribute extensions to the linguistic formalism and linguistic resources for each of the languages of the European Union (EU).

The AUG will allow users to exchange information about the system, including problems, solutions, modifications, enhancements and extensions, as well as research results. The support team will act as a clearing house for AUG contributions that enhance or extend the platform. At the same time, discussions between the EC and users of ALEP and other similar platforms should lead to some convergence of ideas about improving the system and individual components, leading to the definition of further actions.

To obtain further information about ALEP please contact:

Cray Systems-ALEP Support, 151 rue Muguets, L-2167 Luxembourg (Tel: +352 42 77 44; Fax: +352 43 95 94; Email: neil@cray-system.lu (Neil Simpkins)

Information about the AUG can be obtained via email from aug-request@iai.uni-sb.de (Jörg Schütz).

For information about the second AUG workshop and the third ALEP user course see 'Conference Announcements'.

Release of AGFL Home Page

This message announces the release of the World Wide Web Home Page of AGFL (Affix Grammars over a Finite Lattice).

1. AGFL

The AGFL formalism, developed at the University of Nijmegen, The Netherlands, is a formalism in which context free grammars can be described compactly. AGFLs are two level grammars: a first, context free level is augmented with features for expressing agreement between parts of speech. Features are treated as types, and their values may range over the subsets of a given finite set, which explains the acronym Affix Grammars over a Finite Lattice.

AGFL grammars are transformed into a parser by the parser generator OPT. The generated parser is a Recursive Backup parser which computes the values of the affixes on the fly. In this way, fast and efficient parsers can be generated. The formalism is quite simple and limited, and therefore easy to read and write.

AGFL comes with a Grammar WorkBench GWB, supporting the development of grammars and the checking of their consistency.

The AGFL formalism does not require any special hardware. The parser generator OPT runs on regular SPARC-systems and MS-DOS machines (386 or higher) and is relatively small. For instance, the MS-DOS version requires less than 1 Mb harddisk space.

2. AGFL on the Web

AGFL has now been made available to the (computational) linguistic community. We think it can be used by (computational) linguists who are in need of a simple grammar formalism with a fast parser generator, suitable for experimental purposes. Therefore, we have made AGFL available via FTP and, recently, via WWW.

The AGFL Home Page contains information about the AGFL formalism like the AGFL manual, documentation and papers, sample grammars and the latest developments. There is also the possibility to download the software and to register yourself as an AGFL user.

You are invited to take a look at the Home Page and to read the information or to download the software. Please feel free to make use of AGFL and its Home Page; we look forward to hear about your experiences.

We are currently planning an AGFL workshop in June. The latest news about this workshop can also be found on the AGFL Home Page.

The URL of the AGFL Home Page is: http://www.cs.kun.nl/agfl/

The address of the FTP-site is: ftp://hades.cs.kun.nl/pub/agfl/

The organisation of the WWW page should be self-explanatory. The structure of the FTP-site is as follows:

- readme
- DOC : this directory contains a number of relevant papers
- PC386 : this directory contains software for MS-DOS machines and an installation guide.
- SUN4 : this directory contains software for SPARC-stations and an installation guide.

Any questions or remarks with respect to AGFL or the AGFL Home Page can be sent to: www-agfl@cs.kun.nl.

On behalf of the AGFL team, Erik Oltmans (Department of Computer Science, University of Nijmegen, The Netherlands)

The LISA Showcase

[From LISA Forum Newsletter vol.3 no.4, December 1994]

LISA will provide the language processing industry with the most extensive electronic resource of translation tools, services and standards for the localization and internationalization business. The first release will be available to LISA General Assembly members early in 1995.

The **LISA Showcase** will be distributed on CD-ROM providing an extensive reference tool containing detailed information about the products, processes, standards and methodologies for the localization and internationalization business. The heart of the Showcase is a directory and catalogue of localization and language processing tools, supplier profiles, products, and guidelines. It will also contain industry trade journals, standards documents, independent product reviews and European Community localization and software development project reports. The LISA Showcase will give precise and up-to-date information. It will run under Windows on standard hardware with a state-of-art user interface that will cross-reference data for easy information access.

During the past several years the localization and internationalization business has evolved considerably. At the same time the requirement for information has increased dramatically. Users need information about language technology products, services and procedures that can help them make well informed business decisions.

LISA is in a unique position of being able to respond to these requirements. Their international reputation and membership gives them this capability. As the leading resource organization in this business, they have gained the respect and support of the key players in the localization business and have access to the major localization publishers, vendors, product developers and technology users in the industry.

R.R.Donnelley Language Solutions is contributing its database of language technology products and suppliers to the Showcase. The database has been published in print since 1992, as the Language Engineering Directory (the LED) and is the authoritative source of information about commercial activities in the field of language technology. With DLS' cooperation, LISA will be providing LED in CD format as part of the LISA Showcase. The LED will form the base of information about language processing tools and localization service vendors. Under the editorial direction of Rose Lockwood, a leading consultant and localization industry analyst, the LISA Showcase will supplement directory listings with detailed information about products, services and standards which support localization.

Service vendors and language processing software developers will be providing full descriptions of their products and services to help software publishers identify appropriate sources of support in the localization process. The LISA Showcase will also feature detailed information about standards and localization production methods as part of LISA's mandate to promote a more effective localization industry.

The LISA Showcase will be the single most comprehensive information resource for the localization industry. The first release will be available in early 1995 to LISA members and the public.

For more information contact: The LISA, 2bis rue Ad-Fontanel, CH-1227 Carouge, Switzerland (Tel: +41-22-301-5760; Fax: +44-22-301-5761; Email: manobile@divsun.unige.ch)

TERMISTI Research Centre

Thierry J.van Steenberghe

[From ELSNET]

As part of its "Telematics applications" programme, the European Communities have planned a number of linguistic engineering research tasks. Interest at the TERMISTI research centre is focused on projects relating to terminology and terminotics (in particular LE 1.10 and LE 2.1). Our team is seeking to work in partnership with others interested in these fields. The TERMISTI centre is particularly concerned with the application of artificial intelligence to managing multilingual terminological data bases. Our studies have shown that exploiting conceptual networks makes it possible to enrich the data significantly and deal satisfactorily with problems of equivalence. Our work has enabled us to develop a software package that links entries respecting the EURODICAUTOM format by way of a conceptual network and generates defining predicates. This prototype has been tested using microglossaries for highly specialised fields of study, work that has been reported in a number of scientific papers. The interests and skills of the TERMISTI team relate chiefly to:

- developing terminological glossaries for highly specialised fields;
- exchanging terminological data, particularly in S.G.M.L.;
- prepublishing text corpora and determining terminological units;
- devising terminological data base management systems handling conceptual networks;
- modelling multilingual terminology management systems;
- research training in terminology and terminotics.

Centre de recherche TERMISTI, Institut superieur de traducteurs et interpretes (ISTI), 34, rue

Joseph Hazard, B-1180 Bruxelles, Belgium. Tel: +32.2.346.26.41. Fax: +32.2.346.21.34. Email: TERMISTI@INFOBOARD.BE

From: Thierry J. van Steenberghe, RIL, University of Louvain, Department of Computer Science, Place Ste Barbe 2, B-1348 Louvain-la=Neuve, Belgium. (Tel: +32 10 47 3150 [or 2653]; Fax: +32 10 45 0345; Email: tvs@info.ucl.ac.be)

The LOLITA Project at Durham University, UK.

[Extract from job announcement]

Here are a few facts about LOLITA:

- based on a conceptual graph of more than 100k nodes, compatible with WordNet;
- able to perform the fundamental morphological, grammatical, semantical, pragmatical, discourse analysis and generation functions;
- under development for more than 8 years, at present a team of more than 20 researchers working on it;
- mainly written in Haskell, a pure lazy functional language, with high order functions, polymorphic types and type classes (more than 45k lines of code, corresponding to approx 450k lines in an imperative language);
- can handle analysis of real text samples; prototype applications include query, dialogue, template extraction, translation and language tutoring;
- advanced inference capabilities, including multiple inheritance, relevant implication, epistemic reasoning and plausible reasoning (analogy, closed personal world assumption and plausible epistemic);
- processes English and Chinese; Italian and Spanish under development;
- very fast execution times (a parallel version under development);
- applications with Siemens Plessey, Rolls-Royce and other major companies and governmental organisations;
- 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).

WinGlos 2.1: Galician-English-Spanish dictionary

Javier Gomez Guinovart

[From: LINGUIST list 6.131]

WinGlos 2.1 is a trilingual (Galician-English-Spanish) dictionary of computing, in the form of a hypertextual document for Microsoft Windows Help version 3.1, with 500 terminological entries including Galician term, grammatical category of Galician term, English term and Spanish term. Some entries contain a short context of the term, with the aim of elucidating possible ambiguities about the use and meaning of the term. You can access any entry selecting the Galician, English or Spanish term from the alphabetical lists provided by WinGlos.

I have uploaded to SimTel, the Coast to Coast Software Repository (tm), (available by anonymous ftp from the primary mirror site OAK.Oakland.Edu and its mirrors):

ftp://oak.oakland.edu/SimTel/win3/lang/wg21.zip
Galician-English-Spanish computer dictionary

Special requirements: MS-Windows 3.1 (with its Help program WINHELP.EXE).

Further information: Javier Gomez Guinovart, University of Vigo, Spain (Email: uvifejgg@cesga.es)

Japanese Corpora Available

[From LINGUIST list]

(1) spoken Japanese

ATR corpus contains conversations between Japanese speakers through telephone and/or keyboard communications. All conversations are transcribed. Morphological and syntactical tags are given. Corresponding English is given. About half million words are available. The contact address of a distribution coordinator is as follows.

ATR (Advanced Telecommunications Research Institute) International Research Engineering Department Mr. Shohei TAHARA, 2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, Japan. Telephone: +81 774 95 1192. Facsimile: +81 774 95 1179. Email: sho@ctr.atr.co.jp

(2) written Japanese

EDR corpus is available. 28 million sentences are collected from newspapers, magazines and so on. Morphological and syntactical tags are given to about half million sentences.

The contact address of EDR Office changed from April 11, 1995 as follows.

EDR (Japan Electronic Dictionary Research Institute, Ltd.); E-mail: thoth@edr.co.jp. Telephone: +81 3 3851 5521. Facsimile: +81 3 3851 5840

From: Toshiyuki TAKEZAWA takezawa@itl.atr.co.jp, ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan

Linguistic Data Consortium

[From LINGUIST listserver]

The Linguistic Data Consortium (LDC), a nonprofit membership organization affiliated with the University of Pennsylvania, will add about 20 new releases to its 48 existing speech, text, and lexical databases during the current 1995 membership year. The new releases will feature text corpora in six languages, French-English parallel texts, a major telephone speech corpus, and new additions to the existing ARPA speech recognition and spoken language understanding series. Lexicons and large speech corpora in several languages are also in development and scheduled for release in six to nine months.

Consortium membership is annual, with the membership year (MY) running from September to August. Each LDC corpus is identified by the MY of its release, and the annual membership fee purchases a permanent paid-up license to that MY's releases, except that some corpora, owned by others and distributed by LDC, may require a separate user agreement and/or charges.

Members receive one copy of each requested LDC corpus free, and extra copies at a small charge. Nonmember prices are shown in the tables below. Items marked "MO" are for members only, due to restrictions by the copyright owners.

Detailed information about the LDC and a catalog describing its holdings are available via ftp or the World Wide Web (see below); the following is a summary listing of the database titles by year of release.

1993 Releases: TIMIT -- NTIMIT -- Resource Management Complete -- ATIS0 Complete Set -- ATIS2 -- CSR-I (WSJ0) Complete -- SWITCHBOARD -- SWITCHBOARD Credit Card -- TI

46-Word -- TIDIGITS -- Road Rally -- HCRC Map Task Corpus -- ACL/DCI -- TIPSTER Volume 1 -- TIPSTER Volume 2 -- TIPSTER Volume 3
1994 Releases: CSR-II (WSJ1) Complete – CSR-II (WSJ1) Sennheiser – CSR-II (WSJ1) Other -- Air Traffic Control -- SPIDRE -- YOHO Speaker Verification -- OGI Multilanguage Corpus -- OGI Spelled & Spoken Word -- ATIS3 -- BRAMSHILL -- MACROPHONE (American English) -- UN Parallel Text (Complete) -- UN Parallel Text (English) -- UN Parallel Text (French) -- UN Parallel Text (Spanish) -- ECI Multilingual Text -- CELEX Lexical Database -- COMLEX English Syntax Lexicon, Version 0 -- COMLEX Pronouncing Dictionary, Version 0
Planned 1995 Releases: KING Speaker Verification -- Hansard French/English – CSR-III Speech: Dev and Eval Data – CSR-III Text: Language Models – LATINO-40 Spanish Read News Corpus -- WSJCAM0: Cambridge Read News Corpus -- PHONEBOOK: NYNEX Isolated Words -- TRAINS spoken dialogs corpus -- Corpus of Spoken American English-1 -- TIPSTER Volume 4 – Treebank-2 -- Spanish News Text Collection -- North American Business News Text -- Japanese Business News Text -- Mandarin News Text -- French Newspaper Text -- North American Newspaper Text -- Speech Collection Interface SW
Planned 1996 Releases (Tentative): JEIDA Japanese Speech Data -- Corpus of Spoken Amer English-2,3 -- Mandarin Lexicon -- Spanish Lexicon -- Japanese Lexicon -- English Language International News -- Legal Text (500 M words) – POLYPHONE-II (American Spanish) -- Mandarin Telephone Speech -- Japanese Telephone Speech -- Spanish Telephone Speech -- CALLFRIEND Language ID Corpus -- SWITCHBOARD (Revised)
For more information, including membership forms and catalogs:
LDC is at ftp.cis.upenn.edu under /pub/ldc. When accessing by ftp, use "anonymous" as your userid, and your email address for password.

The LDC's World Wide Web Home Page holds the LDC catalog and the "README" files from most of the databases. It can be accessed at URL: ftp://ftp.cis.upenn.edu/pub/ldc_www/ hpage.html

Penn Treebank Project from Linguistic Data Consortium

THE PENN TREEBANK PROJECT
Release 2

The Penn Treebank Project Release 2 CDROM features the new Penn Treebank II bracketing style, which is designed to allow the extraction of simple predicate/argument structure. Over one million words of text are provided with this bracketing applied, along with a complete style manual explaining the bracketing, and new versions of tools for searching and treating bracketed data.

This CDROM also contains all the annotated text material from the earlier Treebank Preliminary Release, including the Brown Corpus. While these materials have not all been converted to the newer bracketing style, they have been cleaned up to remove problems that had appeared in the earlier release.

The contents of Treebank Release 2 are as follows:

* 1 million words of 1989 Wall Street Journal material annotated in Treebank II style.

* A small sample of ATIS-3 material annotated in Treebank II style.

* 300-page style manual for Treebank II bracketing, as well as the part-of-speech tagging guidelines.

* Tools for processing Treebank data, including a new version of tgrep (a tree-searching and manipulation package).

* The contents of the previous Treebank CDROM (Version 0.5), with cleaner versions of

the WSJ, Brown Corpus, and ATIS material (annotated in Treebank I style).
In addition, the Penn Treebank Project will be providing updates, announcements and a discussion forum for users. A file of updates and further information available via anonymous ftp from ftp.cis.upenn.edu, in pub/treebank/doc/update.cd2. This file will also contain pointers to a gradually expanding body of relatively technical suggestions on how to extract certain information from the corpus.

Detailed questions about the corpus may be sent to treebank@unagi.cis.upenn.edu, while questions and requests for obtaining Treebank Release 2 should be sent to ldc@unagi.cis.upenn.edu.

List of some Multingual Text Corpora

[From LINGUIST list]

General addresses::

- LDC material: on ftp.cis.upenn.edu:/pub/ldc
- WWW index at www.ims.uni-stuttgart.de/info/FTPServer.html.
- lexical@nmsu.edu

The INTERSECT (International Sample of English Contrastive Texts) Project at Brighton University began in the Spring of 1994. The aim is to construct and analyse a parallel bilingual corpus of French and English written texts, adding other languages later if resources permit. So far the corpus contains about 5 megabytes of text in each language. The material includes newspaper articles, official documents, instructions for domestic appliances, telecommunications, texts from international organisations, modern fiction, and academic textbooks. Contact: Raphael Salkie, The Language Centre, University of Brighton, Falmer, Brighton, BN1 9PH England. (Email: RMS3@BRIGHTON.AC.UK)

The LINGUA project in Europe is building multilingual corpora for English, French, Greek and some others, for use in language pedagogy. Contact: laurent.romary@loria.fr

The MULTEX project is building tools for multinlingual corpus access, and also a bunch of sample corpora. Contact: veronis@fraix11.univ-aix.fr

A Scandinavian project to build multilingual (English/Swedish/Norwegian/Finnish) parallel corpora. Contact: stig.johansson@iba.uio.no

The European Science Foundation Second Language Acquisition Data Bank (ESFSLDB) contains data of transcribed encounters of untutored language acquisition of adult immigrants. Source languages are Punjabi, Spanish, Finnish, Italian, Turkish, and (Moroccan) Arabic, target languages are English, French, Swedish, Dutch, and German. More details in Perdue, Clive (ed.): Adult Language Acquisition: cross-linguistic perspectives. 2 vols. Cambridge: Cambridge University press 1993.

The Pompeu Fabra University Language Research Institute (IULA) in Barcelona is starting to compile written language corpora. The areas to be covered are law and economics, starting with Catalonian and Spanish languages but to be expanded in the future to English, French and German). Contact: Jorge Vivaldi Palatresi, Universidad Pompeu Fabra, Instituto de Linguistica Aplicada, Rambla Santa Monica 32, 08002 Barcelona, Spain (Email: vivaldi@upf.es)

The European Corpus Initiative Multilingual Corpus I (ECI/MCI) CD was made available in April 1994. ECI was founded to oversee the acquisition and preparation of a large multilingual corpus and supports existing and projected national and international efforts to carefully design, collect and publish large-scale multilingual written and spoken corpora. ECI has produced a multilingual 93 million word corpus covering most of the major European languages, as well as Turkish, Japanese, Russian, Chinese, Malay and more. The primary focus in this effort is on textual material of all kinds, including transcriptions of spoken material. In order to obtain a copy of the ECI/MCI CD, you will need to sign the necessary user agreements. This, together with a copy of the full listing of files on the CD, is obtainable by

(1) anonymous ftp from scott.cogsci.ed.ac.uk/pub/elsnet/eci; or

(2) World Wide Web from http://www.cogsci.ed.ac.uk/elsnet/eci.html.

The University of Surrey (UK) has a number of text corpora in English with their 'shadows' in German, Spanish, Dutch, French and Welsh. The corpora range from 10,000 words to 300,000 words and all the corpora are domain or subject specific.

A parallel German-Norwegian corpus. Contact: Cathrine Fabricius-Hansen, Germanistisk institutt, PB 1004, Blindern, N-0315 Oslo, Norway (Email: c.f.hansen@german.uio.no)

A translation corpus of English and German by Prof. Schmied at the Technical University of Chemnitz-Zwickau. Contact: hildegard.schaeffler@phil.tu-chemnitz.de or josef.schmied
@phil.tu-chemnitz.de

Corpus Linguistics Group (University of Birmingham, UK) Offers Tagging Service on Electronic Mail

[From LINGUIST list]

We are pleased to announce an experimental E-Mail Tagging Service for English texts. The tagging program which is in use at the Corpus Linguistics Group here at Birmingham works stochastically. That means, it calculates the most probable word class in case of ambiguities (if a word can belong to several word classes, like *light*, which can either be a noun, a verb or an adjective, depending on its actual use). Both the probability of the word belonging to a certain word-class and the probability of the word-class occurring at the specified position in the text are taken into account. Since it's probabilistic, there is no 100% correctness, but it is offered as a useful tool rather than a theory of language. We have not formally measured the accuracy of the tagger, but believe it to be quite high.

The program is now publicly accessible by means of an experimental E-Mail Tagging Service. In order to get an English text tagged, just send text to: tagger@clg.bham.ac.uk. The text should not contain any formatting information, as this might lead to undesirable results. The output of the tagging process is sent back to you by email, together with an ID code for later reference.

If you want to get a long text tagged, please split it up into several parts of about 50 KB each, since some mailers cannot cope with huge mails.

If you want to receive the list of tag labels used, just send an empty mail with the subject line "taglist" to the above address.

Since we are using CPU time of one of our workstations, we would naturally like to profit from this enterprise as well. We are always trying to increase our own collection of English text data and so we would like to keep a copy of each text that has been sent to us. So don't send us anything you don't want us to use (eg. if you are not allowed to pass a text to other people).

We do not take any responsibility for damage caused by using the Experimental E-Mail Tagging Service. We also do not guarantee that the results obtained will be correct, even though we will do our best to achieve this. If you notice any errors, we would be glad if you could send a corrected version of your text (ie. with wrong tags replaced by correct ones) to tag-admin@clg.bham.ac.uk. We would then be able to further enhance the quality of the tagger's output.

Corpus Linguistics Group, School of English, The University of Birmingham (WWW-access via http://clg1.bham.ac.uk/; Email: tag-admin@clg.bham.ac.uk)

Software for Lexical Analysis

Paraic Sheridan

[From LINGUIST list]

GERTWOL is the German lexical analysis tool from Lingsoft Inc, who have tools for analysis of many different languages (French and Italian are still under development). Contact: Markku Norberg, Marketing & Sales, Lingsoft, Inc. (Email: norberg@huovinen.lingsoft.fi

The Institut für Maschinelle Sprachverarbeitung (IMS) at the University of Stuttgart have a language-independent tagger that is trained using a hand-tagged corpus and a lexicon. This is available free of charge (incl. an already trained version for German) upon signature of a license agreement. Contact: Helmut Schmid, Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung (IMS), Azenbergstrasse 12, D-70174 Stuttgart, Germany (Email: schmid@ims.uni-stuttgart.de)

INTEX is a finite-state-transducer based morpho-syntactic analyser with dictionaries available for English, French and Italian. It runs on the NextStep operating system. See "INTEX: a corpus processing system", in the proceedings of COLING94. Contact: Max Silberztein (Email: silberz@ladl.jussieu.fr)

The MULTEX project is constructing tools for processing multilingual corpora. Contact: Jean Veronis, Laboratoire Parole et Langage, URA 261 CNRS, Université de Provence, 29 Avenue Robert Schuman, F-13621 Aix-en-Provence Cedex 1 (Email: veronis@grtc.cnrs-mrs.fr,veronis@univ-aix.fr)

Rank Xerox Research have morphological analysis and part-of-speech taggers for many different langauges that are available for commercial and academic purposes. Contact: Daniella Russo (Email: drusso.osbu_north@xerox.com)

INGENIA-Langage Naturel is a French company that markets an analyser for English and French called SYLEX, which provides full syntactic analysis. There is a 2-3 day training course for learning the system. Contact: Patrick Constant, 92 bis, Av Victor Cresson, 92130 Issy Les Moulineaux, France (Email: constant@ingenia.fr)

ARTICLES

The Integration of Linguistic and Domain Specific Knowledge: CAT2 within ANTHEM

O.Streiter and A.Schmidt–Wigger
IAI, Saarbrücken

1 Introduction to ANTHEM

The aim of the LRE project ANTHEM is to develop a prototype of a natural language interface that allows users of Healthcare Information Systems to enter medical diagnostic expressions in Dutch or French (Ceusters et al., 1994a). Within ANTHEM, the CAT2 MTsystem is used to analyse these expressions, translating them into (a) German, Dutch and French and (b) a semantic representation which is then passed to the ANTHEM Expert System for automatic coding in ICD.

The ANTHEM consortium consists of RAMIT Ghent (coordinator), FUNDP Namur, IAI Saarbrücken, CRP-CU Luxembourg, the University of Liège, Datasoft Management nv Oostende and the Military Hospital Brussels. CAT2 is a unification-based machine translation system developed at IAI Saarbrücken; the most up-to-date description can be found in Sharp & Streiter (1995).

A first prototype of ANTHEM running on a Unix Workstation was realized in 1994 (cf. Ceusters et al. 1994b). Currently the lexical coverage and the facilities of the system are being extended in order to allow for an application in a real life medical setting. The portation to DOS is under way.

2 The Construction of an Interlingua

It has been known for a long time in MT theory that the interlingual approach is the most promising in multilingual systems. Since in ANTHEM at least three natural languages and one language-independent representation are involved, the interlingual approach seems to be the most natural to follow. In order to circumvent the main problem – the preliminary construction of an interlingual classification of concepts for the whole domain (cf. Arnold & Sadler 1992) – an existing classification in Medicine, the Systematized Nomenclature of Medicine (SNOMED) (cf. Côté et al. 1993), has been adopted for this purpose . For every language involved the coupling of words to concepts is done in the CAT2 lexicons, in which every lexical entry contains the associated SNOMED code (e.g.*snomed='M-12000'*) apart from the lemma (e.g. *lemma=fracture*) and its grammatical description.

In order to express generalizations about the concepts, they are regrouped into 27 classes (Semantic Types). Some classes are taken from the dimensions used in SNOMED (e.g. *topography, morphology, function*), others such as *living_object* and *chemical* are identified by the paradigmatic relations which they maintain (for a complete description see Ceusters et al. 1994b).

In order to describe the combination of Semantic Types a set of Semantic Roles has been developed. Following linguistic models, each possible head of a structure was assigned an argument frame, i.e. a set of arguments to which the head can assign a Semantic Role, plus the restrictions on the Semantic Type in order to control the access to the argument slot. The set of Semantic Types, Roles and Restrictions is called the 'ANTHEM Semantic Model' which represents the interlingual domain specific knowledge on which the analysis in CAT2 is based.

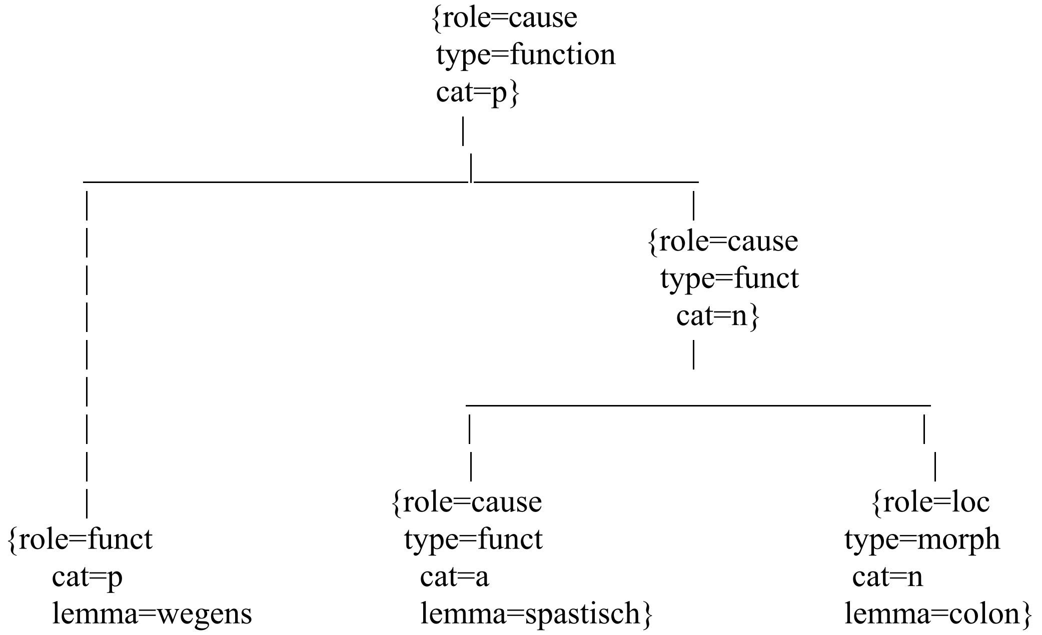
3 Implementation in CAT2

3.1 Basic Schemes of Composition

The syntactic and semantic analysis is carried out with a limited set of schemes of composition which represent the structures by which larger expressions are formed. The generic parser implemented in CAT2 uses these schemes of composition to build valid tree structures. The schemes are responsible for the construction of (1) head–argument structures, (2) functional projections, (3) coordinated structures and (4) multi word units.

The main scheme of composition, the head–argument structure, takes into account only semantic (i.e. domain specific) information and percolates the semantic properties of the semantic head (i.e. the part that functions as predicate) to the mother node. Every tree structure built by this scheme is submitted to a syntactic verification rule which may filter out illegitimate structures. This rule tests the position and inflection of adjectives and adjacency constraints (e.g. the position of of-phrases and genitive phrases). Syntactic information is furthermore used in preference rules for disambiguation of the semantic analysis. Examples are the embedding of topologies where the third localises the second and not the first, or the preference of a nominal over an adjectival semantic head for some type distributions, in which, given the ANTHEM Semantic Model, both could function as semantic head.

Moreover the verification rule identifies the syntactical head (i.e. the daughter which shares its syntactic properties with the mother node). Through this `split' analysis, in some structures the mother node receives its properties from different daughters. The necessity to calculate separately the semantic and the syntactic head becomes obvious when, for example, a preposition selects the semantic properties of one daughter and the syntactic properties of the other. In the following example the Dutch preposition wegens ('due to') semantically selects the features *role=cause, type=function* which comes from the adjective, but the syntactic properties (e.g. *category, case*) of the nominal head.



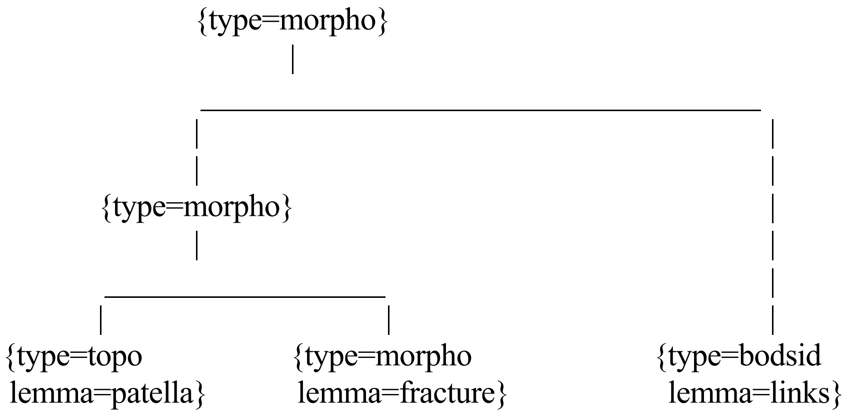
Functional categories (e.g. prepositions and determiners) are added through the scheme of functional projection, with the functional category (e.g. the preposition) as syntactic head and the lexical category (e.g. the noun) as semantic head of that structure.

Coordination of constituents (e.g. VIRAAL + BACTERIEEL) is accounted for by a trinary scheme of composition. The syntactic head of the coordination is the coordinator as the syntactic properties of the coordination may change with respect to that of the coordinated substructures (e.g. in number). Both coordinated substructures function as semantic head sharing their Semantic Roles and Types with the mother node. By recursive application of this scheme, coordinations of any complexity can be analysed.

A multi-word unit (MWU) is a fixed sequence of words which noncompositionally refer to one concept of the semantic model (e.g.*Vitamin A*). Such MWUs are built by the fourth scheme of composition which connects the syntactic head (e.g. *Vitamin*) with the non-heads (e.g.*A*), recognized by their dummy type *type=mwu*.

3.2 Discontinuous Structures

The medical diagnostic expression may contain discontinuous structures, where the argument is separated from its head by a third element. In the following example the *bodyside*–marker LINKS ('left') has undergone a rightward movement so that it is syntactically connected to the projection of fracture ('fracture') but refers semantically to PATELLA**.**



The implementation of discontinuous structures is based on the *slash* principle as described in Gazdar et al. 1985). If the *bodyside* slot of a *morphology* is not filled when its projections becomes the semantic non-head of a second structure, the empty slot is passed within the *slash* feature onto the semantic head.

3.3 Compounds

Compounding is a morphological process, which is equivalent to syntactic means to combine concepts. Dutch and German compounds are analysed into their parts: the head of the compound (the rightmost component) contains in its feature bundle a complete description of the non-head, as shown for the Dutch compound LONGONTSTEKING**.**

lemma=ontsteking, type=morpho, head={cat=n}
non_head={lemma=long, type=topo, head={cat=n}}

A verification rule determines the semantic nature of the compound through the unification of the non-head with one of the argument slots of the head. For the purpose of translation, head and non-head will be broken into two separate tree structures so that the translation is a compositional one, where each part has its equivalent in the target language (SPIER ↔ MUSCLE, SPASME ↔ SPASME).

PARAVERTEBRALE SPIERSPASMEN
('para–vertebral muscle spasms')
SPASMES PARAVERTEBRAUX DANS LE MUSCLE
('spasms para–vertebral in the muscle')

In medical sublanguage, however, compounds like MWUs often refer as a whole to one concept only. This is accounted for by a seperated lexical entry with one concept code for this concept, blocking the analytical translation process.

TENNISELLEBOGEN	('tennis elbow')
* COUDE DE TENNIS	('elbow of tennis')
EPICOINDYLITE	('epicondylitis')

3.4 Compounds and Extraposition

In the same way as the phrasal structures, compounds allow for the extraposition of the *bodyside* and thereby for the appearance of discontinuous structures, where the *bodyside* refers to the non-head of the compound. This structure is accounted for by the same *slash* mechanism: If a *morphology* is appearing as a non-head of a compound, the *bodyside* slot is transmitted within the *slash* feature to the head, from where the *bodyside* marker can be bound.

ENKELDISTORSIE LI	('ankle-contortion left')
POLSONTSTEKING RECHTS	('wrist-inflammation right')

3.5 Lexical Functions

The concept of *Lexical Functions* has been developed by Mel'čuk in the framework of his Meaning ↔ Text Model (cf. Mel'čuk 1974). The essence of this notion is that word A selects a second word B in order to realize a special meaning related to A which is called the lexical function LF. Assume A to be smoker . In order to form the high degree of it, which is not morphologically possible in English, A selects B = HEAVY as its modifier in order to realize through the expression HEAVY SMOKER the high degree of smoker. Lexical Functions merit a special treatment in MT since A but not B can be translated literally.

In ANTHEM we find lexical functions within the combination of *severity* and *function*. In many cases the *function* selects one special *severity* operator for the Magnifier and another, not necessarily related word, for the Minifier; examples are taken from the German corpus.

SCHWERE/LEICHTE VERBRENNUNG	('heavy/light burn')
HOHES/LEICHTES FIEBER	('high/light fever')
STARKE/SCHWACHE SCHMERZEN	('strong/weak pain')
STARKE/LEICHTE RÖTUNG	('strong/light reddening')

The different realizations of the high degree (i.e. the words schwer/hoch/stark) and the low degree (i.e. leicht/schwach) receive the same interlingual representation. Which lexeme is to be chosen for its realization is determined in the lexicon for every *function*.

4 Conclusion

In the preceding discussion we have shown how the CAT2 system analyses the natural language input of ANTHEM based on an interlingual domain specific semantics and syntactic and lexical restrictions. The main difficulties with the analysis and translation of this input arise from the mismatch of syntactic and semantic principles (e.g. principles of projection and principlesof lexical selection). As a consequence, the semantic and syntactic constraints which represent the domain specific and the language specific knowledge respectively, apply in sequence in order to assure an efficient and correct analysis and translation.

5 Bibliography

Arnold, D. and L. Sadler (1992): Unification and machine translation. *Meta*, 37(4), 1992: 657--680

Ceusters, W., G. Deville, E. Herbigniaux, P. Mousel, O. Streiter, and G. Thienpont (1994b): *The ANTHEM Prototype.* Working paper 31, 1994, IAI, Martin-Luther-Straße 14, 66111Saarbrücken, BRD.

Ceusters, W., G. Deville, O. Streiter, E. Herbigniaux, and J. Devlies (1994a): *A computational linguistic approach to semantic modeling in medicine.* in Belgo-Dutch Congress on Medical Informatics '94, Veldhoven, 1994: 311-319.

Côté, R.-A., D.–J. Rothwell, R.–S.Beckett, and J–L. Palotay, editors (1993): *Developing a standard data structure for the systematized Nomenclature of Human and Veterinary Medicine. SNOMED International. Introduction.* College of American pathologists & American Veterinary Medical Association, 1993.

Gazdar, G., E. Klein, G. Pullum, and I. Sag (1985): *Generalized Phrase Structure Grammar.* Oxford: Blackwell, 1985.

Mel'čuk, I.A. (1974): *Opyt teorii linguističeskich modelej Smysl ↔ Tekst.* Semantika, sintaksis. Izdatel'stvo Nauka, Moskva, 1974.

Sharp, R. and O. Streiter (1995): *Applications in multilingual machine Translation.* in: Proceedings of The Third International Conference and Exhibition on Practical

Applications of Prolog, Paris, 4th-7th April, 1995.

FROM THE ARCHIVES...

"The whisky was invisible", or Persistent myths of MT

John Hutchins

Scarcely a month goes by without somebody repeating the story of the MT system which translated the Biblical saying "The spirit is willing, but the flesh is weak" into Russian, which was then translated back as "The whisky is strong, but the meat is rotten". If this is not quite as you remember it then that is understandable. Perhaps you heard that the back translation was "The vodka is strong but the steak is lousy". Or maybe you heard that the language was not Russian, but German... or Japanese... or Chinese...

Often the story is told to show how poor the older approaches to MT can be, as in this press release promoting the new LMT system for PCs, the Personal PT, in October 1994:

> Die Wort-für-Wort Übersetzungssysteme sind einfach damit überfordert, die Komplexität der menschlichen Sprache auch nur annäherend zu verstehen. So wird der Biblespruch "The spirit is willing, but the flesh is weak" (Markus 13.4) sinngemäß ins Russische übersetzt mit: "Der Whisky ist stark, aber das Fleisch ist faul."
>
> [It is simply beyond word-for-word translation systems to understand the complexity of human language even approximately. Thus the Biblical saying 'The spirit is willing, but the flesh is weak' is translated into Russian as the equivalent of 'The whisky is strong, but the meat is rotten'.]

Workers in artificial intelligence have also used the example. Elaine Rich (**Artificial intelligence.** New York: McGraw-Hill, 1984) describes one of the problems of MT systems which do not understand text and translate 'meaning' as an inability to deal with idioms (p.341):

> An idiom in the source language must be recognized and not translated directly into the target language. A classic example of the failure to do this is illustrated by the following pair of sentences. The first was translated into Russian, and the result was then translated back to English, giving the second sentence:
>
> 1. The spirit is willing but the flesh is weak
> 2. The vodka is good but the meat is rotten.

As these extracts show, MT and AI researchers have cited this 'howler' to illustrate problems of ambiguity and lexical selection which are supposedly typical of older 'word-for-word' or 'direct translation' systems and which their own systems are presumably able to deal with successfully.

There is of course no hint in these quotations that the example may not be anything other than a genuine output from an MT system. This is characteristic; the example is used to illustrate a weakness of some unnamed 'earlier' system.

Nearly always when the story is repeated there is no suggestion that it might be apocryphal. Admittedly, some writers have shown doubts. Isidore Pinchuk (**Scientific and technical translation**. London: Deutsch, 1977) in his chapter on machine translation writes (p.241):

> It has often been said that a computer is an idiot, and many examples of its imbecility (probably apocryphal) have been given. 'The ghost is a volunteer but the meat is tender', an alleged computer translation of *der Geist ist willig, aber das Fleisch ist schwach*, underlines the fundamental problems inherent in

MT...

The initial source for many of the recent versions of this MT howler may well be an article in the influential and widely read **Harper's Magazine**. In August 1962, John A. Kouwenhaven wrote an article "The trouble with translation", which included the following paragraphs:

> Our own attempts to communicate with the Russians in their language may be no more successful. Thanks to Robert E. Alexander, the architect, I can pass along this cheering bit of news. According to Colonel Vernon Walters, President Eisenhower's official interpreter, some electronic engineers invented an automatic translating machine into which they fed 1,500 words of Basic English and their Russian equivalent, claiming that it would translate instantly without the risk of human error. In the first test they asked it to translate the simple phrase: "Out of sight, out of mind." Gears spun, lights blinked, and the machine typed out in Russian: "Invisible Idiot."
>
> On the theory that the machine would make a better showing with a less epigrammatic passage, they fed it the scriptural saying: "The spirit is willing, but the flesh is weak." The machine instantly translated it, and came up with "The liquor is holding out all right, but the meat has spoiled."

This account has all the appearance of genuineness. It was well known at the time that much research on MT was going on in the United States and that Russian was the main language of interest. It is, therefore, not surprising that many believed it to be true and so repeated it.

However, there is no evidence of a system of the kind described was in existence in the early 1960s; after all, the Americans were concerned not with translating into Russian but with translating technical and military documents from Russian. And why should the system be restricted to Basic English? It would have taken just a little knowledge of Russian, or indeed of any language other than English, to throw doubts on its authenticity. While *flesh* could well be back-translated as *meat*, and *spirit* could conceivably come out as *liquor*, how could anyone believe that *willing* was translated as the equivalent of *holding out* and *weak* as *spoiled*?

Fortunately we do not have to tax our brains to work out how this MT program might have produced such 'howlers', since both were already known some years before.

In February 1958, A.G.Readett gave a brief account of a lecture (**Linguists' Review**, NS vol.1 pt.2, p.27-28) on progress in research on a French-English translation system at Birkbeck College. At the end of his paper there appeared the following paragraph:

> **Apocryphal**
>
> A Firm experimenting with an electronic brain designed to translate English into Russian fed it with the words: "The spirit is willing but the flesh is weak."
>
> The machine responded with a sentence in Russian characters which was handed to an expert linguist.
>
> "It says," he reported, "that the whisky is agreeable but the meat has gone bad."

But why was this particular saying chosen? As it happens two years earlier in April 1956, a lecture was given at the Institute of Electrical Engineers on the "computer in a non-arithmetical role", and MT was mentioned as one potential application. In the subsequent discussion, one of the participants E.H.Ullrich may have been the unwitting originator:

> Mechanical translation will surely come, and I welcome the attempts at it now being made. I feel, however, that most of the workers in this field underestimate by a factor of ten the difficulty of producing a useful and truthful translation as opposed to a novelty for amusement only. They appear to think that dictionaries and grammars together contain substantially all that is required for the purpose. In serious matters this is usually not so. Before the war, I lived for a number of years in Paris and found the standard of translation in the Press poor.

> Perhaps the popular Press is the most attractive outlet for mechanical translations, because it does not really matter whether these are right or wrong, and amusing versions such as 'the ghost wills but the meat is feeble' might make mechanical translation into a daily feature as indispensable as the cross-word puzzle. ...

It is surely ironic that a joke by journalists about incompetent human translators should be used, in all seriousness, to show how poor computers can be in comparison with human translators.

Interestingly, the talks which gave rise to the apocryphal story and to Ullrich's comment were both given by the pioneer MT researcher Andrew Booth, who himself revealed the antiquity of that other MT perennial "Invisible idiot". In a review of a report of Lamb's MT system for Chinese-English translation (Nature, 200 (2 November 1963): 392-393) entitled *Invisible lunatic*, he wrote:

> The second of the objectives, that of assembling all the Chinese characters in modern use, is more difficult an assessment and suggested the title of this review. There is an old (and probably apocryphal) story that a certain computer was asked to translate "Out of sight, out of mind" into Chinese and that a second machine later preformed the reverse process, producing as a result the words of the title of this review. The point of the story is that, although telegraphic entries nos. 4035, 4045 and 4082 are associated with madness and imbecility, neither the word lunatic nor invisible seems to appear in the list...

The late Margaret Masterman (another MT pioneer) used to say that this particular mistranslation was frequently to be found in elementary textbooks for learning Chinese as a warning against just such a careless use of character dictionaries that Booth was describing.

Like the Biblical saying, this story too persists, appearing sometimes in most unexpected places, as illustrated by the following extract from a British provincial newspaper (**Eastern Daily Press**, December 19, 1987):

> Computers will never take over... Citing evidence for this re-assuring proposition, Benedict Cadbury, factory manager of UB Frozen Foods, Fakenham, quoted to his audience a story of how a computer was programmed to translate English into Japanese. "Out of sight, out of mind" was the task set the machine. Out came an impressive Japanese print-out, followed by the acid test, a re-translation into English. The result was a two-word precis: "Invisible idiot".

If nothing else, this should serve as a warning not to believe everything read in newspapers. MT has suffered repeatedly from misleading journalism - examples were given in MTNI#8 of the wild exaggerations in reports of the IBM-Georgetown experiments. It may be difficult for the MT community to combat such stories since they have undoubted humour and memorableness, and since they are sometimes believed even by those involved in the field. However, there are surely more than enough true MT 'howlers' which could be cited - such as the translation of *les enfants et les femmes enceintes* as *pregnant children and women* - without the need to resort to those which are known to be apocryphal.

[The editor will welcome any further examples of MT howlers, and any explanations of their origins, for publication in future issues of MTNI.]

Translation strategy

Michael Blekhman

When writing about my system PARS [e.g. in this issue], I wanted to make use of the "almost-classic" definition of the three translation approaches: direct, transfer-based, and interlingua-based. However, I perceived that it was hardly possible to use this definition practically as it was very hard to draw a demarcation line between the above three approaches. The first reason was that the champions of this definition consider what they call "direct translation" quite fruitless, while, on the other side, systems translating "directly" are sold and, what is more important, bought throughout the world, giving their developers honestly earned profits, the latter being sometimes rather high. So, I came to the conclusion that a different kind of terminology was to be suggested.

Describing PARS translation philosophy, I have to point out that, in PARS dictionaries, grammatical and semantic information is attributed to the Russian part of the word-entry only. That is why the translation program cannot analyze the source text without analyzing the intermediate product which is the first approach to the target text. PARS makes a word-for-word translation, and brushes it up intensively, making it look as natural as possible. That is why we call our approach FTA: *"first-translate-then-analyze"*. Generally speaking, FTA is usually resorted to if system developers do not want to view the sentence as a single structural entity, considering it as a linear sequence of lexical units and regarding syntactic and semantic relations merely for disambiguation purposes.

On the contrary, a system may first analyze the source text, and then translate it, using the results of this analysis, thus working according to the FAT (*"first-analyze-then-translate"*) principle. Traditionally, the FAT-type systems consider the sentence as a syntactic (or even semantic-syntactic) unit, the basic idea being that the more information you use in your analysis, the better results you will obtain. What is not taken into consideration, however, is that "redundancy errors" are probable in such cases, i.e. "when she (translation algorithm) was good, she was very, very good, but when she was bad, she was horrid": inevitable mistakes in the analysis of as complicated an entity as a sentence will cause translation mistakes. The situation is very well known to practical developers of linguistic information systems, who constantly face the "noise/completeness" dilemma. From time to time, we come across the typical situation: too much analysis causes poorer translation quality than no analysis at all. Our opponents may contradict that "too much analysis" means "too little analysis", but have you ever seen enough analysis in real-life MT systems?

PARS bears on dozens of rules to analyze the source text and synthesize the target one, some of the rules being rather sophisticated, such as disambiguation of "-ed" forms for English-Russian translation. However, it does not dare to view the sentences as structural units. The program only analyzes a word if it is grammatically ambiguous. At the same time, the set of rules is constantly extended in the system "growing" process: we analyze translation results, and if a mistake is typical, that is a certain ambiguity type is come across regularly, we try to develop a rule to eliminate the ambiguity.

PUBLICATIONS: announced and received

International Quantitative Linguistics Association

[From LINGUIST list]
In the last years, *Quantitative Linguistics* has undergone a rapid and promising development, with respect to both theory and application, and quantitative methods are constantly gaining importance in all branches of language and text research. The quantitative approach to language opens up important and exciting theoretical perspectives, as well as solutions for a wide range of practical problems, by introducing into linguistics the methods and models of advanced scientific disciplines such as the natural sciences, economics, and psychology. Quantitative mathematical methods (probability theory, stochastic processes, differential and difference equations, fuzzy

logic and set theory, function theory etc.) are being applied to all aspects of language and text phenomena, including the areas of psycholinguistics, sociolinguistics, dialectology, pragmatics, etc., and on all levels of linguistic analysis. In applied linguistic disciplines, the quantitative approach is constantly gaining interest, e.g. in the fields of natural language processing, machine translation, language teaching, documentation and information retrieval.

In view of the growing number of scientists involved in, and research on, theoretical and applied aspects of QL, an international forum for information and cooperation in this field seemed highly desirable.

Therefore, on the occasion of the Second International Conference on Quantitative Linguistics, QUALICO-2, which was held in Moscow, Russia, at Moscow State University, in September 1994, the International Quantitative Linguistics Association (IQLA) was founded.

At present, the IQLA Council consists of: R. Koehler, University of Trier, Trier, Germany (President); A. Polikarpov, Moscow State University, Moscow, Russia (Vice-President); S. Embleton, York University, Toronto, Canada (Treasurer); P. Schmidt, University of Trier, Trier, Germany (Secretary-General); N. Darchuk, O.O. Potebnya Institute of Linguistics, Kiev, Ukraine; Ju. Krylov, St. Petersburg Electrotechnical University, St. Petersburg, Russia; A. Šhajkevih, Institute of Russian Language, Russian Academy of Sciences, Moscow, Russia; G. Sil'nickij, Smolensk Pedagogical Institute, Smolensk, Russia (Council Members at large).

IQLA membership includes subscription to the official organ of the Association: *The Journal of Quantitative Linguistics*. The actual IQLA annual fees are:

60 US $ (including JQL) for non-student members
20 US $ for students
200 US $ (including JQL) for institutions
(special fees for former Socialist countries and developing countries)

For further information on the IQLA, please contact: IQLA, Universität Trier, FB II, LDV, D-54286 Trier, Germany (Email: koehler@ldv01.Uni-Trier.de)

Contributions to JQL should be sent to: Prof. Dr. Reinhard Koehler (Editor JQL), Universität Trier, FB II, LDV, D-54286 Trier, Germany (Email: koehler@ldv01.Uni-Trier.de); or to: The Journal of Quantitative Linguistics, Editorial Office, Swets & Zeitlinger, P.O. Box 825, NL-2160 SZ Lisse, The Netherlands; or: The Journal of Quantitative Linguistics, Editorial Office, Swets & Zeitlinger/SPS, P.O. Box 613, Royersford, PA 19468, USA.

Bibliography on Connectionist-Symbolic Integration

Ron Sun

[From LINGUIST list]

In relation to the IJCAI Workshop on Connectionist-Symbolic Integration: From Unified to Hybrid Approaches, to be held at IJCAI'95 Montreal, Canada, August 19-20, 1995, I would like to update a bibliography of work on connectionist-symbolic integration.

About two years ago, I compiled a bibliography on the above topic (available in Neuroprose). However, since then, there has been a considerable amount of new developments that need to be collected and categorized. Therefore, I want to update (and re-compile) that bibliography.

Please send me any of the following:

-- New publications since Spring 1993
-- Earlier publications that were inadvertently omitted in the current bibliography
-- Lists of your own publications in this area, preferably annotated (if they are not already in the bibliography).

My e-mail address is: rsun@cs.ua.edu

If you have hardcopies that you can send me, here is my address: Dr. Ron Sun, Department of Computer Science, The University of Alabama, Tuscaloosa, AL 35487 (Tel:(205) 348-6363

The previous bibliography (36 pages) on connectionist models with symbolic processing is available in neuroprose. To get a copy of the bibliography, use FTP as follows:

```
unix> ftp archive.cis.ohio-state.edu (or 128.146.8.52)
Name: anonymous
Password:
ftp> cd pub/neuroprose
ftp> binary
ftp> get sun.nn-sp-bib.ps.Z
ftp> quit
unix> uncompress sun.nn-sp-bib.ps.Z
unix> lpr sun.nn-sp-bib.ps (or however you print postscript)
```

A clean-up version of the bibliography is published in the book: Ron Sun and Larry Bookman. (eds.) Computational architectures integrating neural and symbolic processes (Kluwer, 1994.)

Multilingual PC Directory on WWW

Ian Tresman

[From LINGUIST listserver]

I am pleased to announce new availability of The Multilingual PC Directory, the source guide to multilingual and foreign language software for IBM PCs and compatibles.

1 The full text is now on the World Wide Web at the site:
http://www.knowledge.co.uk/xxx/

2 It's also in Windows Help File format at the site:
ftp://vespucci.iquest.com/tatro-enterprises/insoft-l.arc/classifieds/
babel.zip: 1K description.
mpcdir.zip: 625K full-text WinHelp format.

3 A 256-page book is also available.

You'll find a 1200 word description at these sites, or on request from me at 72240.3447@compuserve.com

Linguistic Databases proceedings

The proceedings of the conference on Linguistic Databases, 23-24 March 1995 held at Groningen's Centre for the Behavioral and Cognitive Neurosciences are not due to appear until the end of this year. Abstracts are available at: http://www.let.rug.nl/~langdb

Computational Linguistics and Chinese Language Processing

Computational Linguistics and Chinese Language Processing invites submission of original research papers in the area of computational linguistics in general and Chinese (natural) language processing in particular. Contributions can be written either in English or Chinese. English will be the primary language of this journal for its international readership. A 600-word extended abstract is required of all Chinese contributions. Submissions are welcomed in the following three categories:

Papers: Submissions in this category should report significant new research results in computational linguistics or new system implementation involving significant theoretical and/or technological innovation. There is no strict length limitation on this category but it is suggested that manuscripts not exceed 40 double-spaced pages.

Survey Papers: Submissions in this category are either invited by the editorial board or voluntary. They should offer a critical overview of either 1) the state of arts of a certain sub-field of computational linguistics or Chinese language processing, or 2) of existing systems and/or technologies for a particular natural language application. Page limitation of this category is the same as above.

Book/Thesis Review: Submissions in this category should not exceed one journal page; i.e. 600 English words/Chinese characters inclusive of titles. The theses reviewed must be completed within one year before the time of submission. An extended abstract can also be submitted by the author if it is endorsed by an editorial board member.

All contributions will be anonymously reviewed by at least two reviewers, with exception of the reviews, which will be reviewed by only one reviewer.

Electronic submissions are required. Submissions should be in postscript files conforming to the Journal of Computational Linguistics style. Submissions should be sent to the editor-in-chief at clp@hp.iis.sinica.edu.tw

Supplementary hardcopy submission is accepted at Computational Linguistics and Chinese Language Processing, Professor Keh-jiann Chen, Institute of Information Science, Academia Sinica, Nankang, Taipei, Taiwan 115

Computational Linguistics and Chinese Language Processing will be published both electronically and in hardcopy. Any paper accepted by the editorial board will be put at the WWW site and considered published. Hardcopy publication will be twice annually in June and December and will be the compilation of the papers electronically published in previous six months. The first issue will appear in June 1996.

Computational Linguistics and Chinese Language Processing is published by the Computational Linguistics Society of R.O.C. (ROCLING).

Hermes - Journal of Linguistics, vol. 13, 1994

Karen M. Lauridsen

Table of contents:

Thematic section: Corpus Linguistics

Lauridsen, K. & O. Lauridsen: Editorial

Asmussen, J. "The Text Corpus and Corpus Retrieval System of The Danish Dictionary"

Barkema, H. "The Idiomatic, Syntactic and Collocational Characteristics of Received NPs: some basic statistics"

Blom, B. "Therminometry: how termish is a term?"

Esser, Jürgen "Medium-transferability and corpora: Remarks from the consumer end of corpus linguistics"

Haslerud, V. & A.-B. Steenström "COLT: Mark-up and Trends"

Kirke, J.M. "Using VARBRUL for Studying Modal Auxiliary Verbs"

Knowles, G. "Annotating large speech corpora: building on the experience of MARSEC"

Leitner, G. ""Begin" and "start" in British, American and Indian English"

Mosey, B. "A study on nuclei in the SEC - report from the public speaking project in Lund"

Nevalainen, T. & H. Raumolin-Brunberg "Sociolinguistics and Language History: The Helsinki Corpus of Early English Legal Texts"

Norre Nielsen, J. & A. Wichmann "A Frequency Analysis of Selected Modal Expressions in German and English Legal Texts"

Paradis, C. "Compromiser - a notional paradigm"

Schmied, J. "Translation and Cognitive Structures"

Souter, C., G. Churcher, J. Hayes, J. Hughes & S. Johnson "Natural Language Identification using Corpus-Based Models"

Wichmann, A. "FO troughs and prosodic phrasing: an investigation into the linguistic information contained in a speaker's base-line when reading"

Tribune des Industries de la Langue: special issue

[From ELSNET]

A special issue of the **Electronic Information Systems and Language Industry Tribune** (90 pages) devoted to the various aspects of the translation –oriented summit is going to be published at the end of April 1995.

This issue can be ordered at Ofil, 2 rue Abel 75012 PARIS (Fax : + 33 (1) 40 02 03 50)

Price: for Academic organisations: US$ 110; for Industrial organisations: US$ 180

For more information please contact Prof. André Abbou at OFIL.

Contents :

Programmes-European Union Commission: - How to get out of the 20th century? (André Abbou and Virginie Boutin) - Setting up the global information society (André Abbou and Virginie Boutin) - A credible and efficient multilingual strategy to be raised (André Abbou)

French-speaking countries policy (Olivier Lesage): - France : Translation policy, language policy - ACCT/Aupelf-Uref : French speaking area gets organised into structures - Canada : A translation policy to maintain social cohesion.

MT-CAT perspectives – All paths lead to CAT: - Present trends (André Abbou) - New semantics for MT (Gaston Gross) - From last to future years (Jean-Marie Lange) - AlethTrad, an aid to translation (Jose Vega) - Personal CAT for monolingual translator (Hervé Blanchon) - Constructive process theory (André Abbou and Pascale Morey)

Markets: - The world-wide global translation market and its components: scientific and technical translation market, C.A.T and translation tools markets (André Abbou)

Strategies: - New trends in Industrial strategy and marketing strategy (André Abbou)

Products: - From bridge to bridge (André Abbou)

Evaluations – From an evaluation to the other (André Abbou, in collaboration with Virginie Boutin and Pascale Morey): - Eagles, or the long walk - Are you speaking about methodologies? - To measure text-system interaction - A good guide to TOEFL - ARPA MT: from methods to perspectives.

Projects: Year 2000's stakes - Global information society - A perspective or a mirage? (André Abbou): - France: Can you see teleservices coming? - Internet networks: Vadememecum - United States: Doing the splits - European Union: Challenges and tracks - Germany: Spreading one's network - Japan: From project to reality -Canada: Jumping at the opportunity - Quebec: Building as big as possible - Electronic information – On-line 1994 (Stephane Chaudiron)

Poles: - Japan: A translation policy, even in a critical period (André Abbou) - Central & Eastern Europe: Research in man/machine interaction in Ukrainia (Françoise Noël and D. Teil) - Arabic countries: Arabic language facing language industry (Christian Fluhr)

Events (Astrid Gillard)

Two reports from OVUM coming

Ovum Ltd (London) has announced the imminent publication of two reports of considerable interest to the MT community. The first report **Globalisation: Creating New Markets with Translation Technology** will be published in May 1995 at £1195 or US$2220. It is intended to be "the definitive guide to the processes and technologies which produce local language products and documentation", and will contain "ten instructive case studies of leading users of globalisation products, as well as eleven comprehensive profiles of vendors providing integrated globalisation tools and services." The authors are Rose Lockwood, Jean Leston, and Laurent Lachal.

The second report **Ovum Evaluates: Translation Technology Products** is to be published in June or July 1995, priced £995 or US$1850. It is a guide to the market, which "reviews the tasks and applications for which each type of product is suitable, compares the different types of tool on offer, presents an easy-to-use framework for evaluating and understanding the products." and "provides an authoritative and independent source of information for companies evaluating the technical and commercial issues in using translation technology." It concentrates on the European and North American markets and includes evaluations of the MT products: DP/Translator, Globalink, LMT, Logos, Metal, Systran; the translator workbenches: Eurolang Optimizer, Star Transit, Trados, TranslationManager, XL8; and the terminology tools: MTX, MTX Reference, Translexis. The authors of the report are June Mason, Adriane Rinsche, and Rose Lockwood.

For more information contact: Ovum Ltd., 1 Mortimer Street, London W1N 7RH. Tel: +44 171 255 2670; Fax: +44 171 255 1995; Email: info@ovum.mhs.compuserve.com (Internet), or MHS:INFO@OVUM (Compuserve).

The Translator

The first issue of a new journal has appeared entitled *The Translator*, published by St.Jerome Publishing (Manchester, U.K.). It intends to publish "articles on a variety of issues related to translation and interpreting as acts of intercultural communication, ...to cover all types of translation, whether written or oral, including activities such as literary and commercial translation, various forms of oral interpreting, dubbing, voice-overs, subtitling, translation for the stage, and such under-researched areas as sign language interpreting and community interpreting." In this first issue there are articles on copyright (Lawrence Venuti), interpreting (Ruth Morris, Sarah Williams), and semantic compensation in translation (Keith Harvey), a reevaluation of Mounin's classic work on translation theory (Juan Sager), and a series of book reviews.

New Books

Kitano, Hiroaki: **Speech-to-speech translation: a massively parallel memory-based approach**. Boston/Dordrecht/London: Kluwer Academic Publishers,1994. xvii,192pp. ISBN: 0-7923-9425-9.

This is the most complete account available of the research conducted at the Center for Machine Translation of Carnegie Mellon University on speech translation. The author was the principal investigator on the ΦDMDIALOG project and on subsequent projects known as DMSNAP, ASTRAL and MEMOIR - all based on massively parallel computational models. The distinctive features of the ΦDMDIALOG research were the exploration of the memory-based approach to natural language processing, predictive parsing, cost-based ambiguity resolution, and generation simultaneous with analysis. Several experiments were conducted using the ATR corpus of telephone dialogues. The subsequent DMSNAP project experimented with a version implemented on the parallel processor SNAP (Semantic Network Array Processor), and the following ASTRAL project implemented the approach on the IXM2 associative memory processor (developed at the Electrotechnical Laboratory in Tokyo), with impressive improvements in processing speeds. Finally, in the MEMOIR project a somewhat different architecture was explored in which rule-based processing was employed to monitor output from memory-based processes.

While the book shows its origins as only slightly revised independently published articles and reports, and contains rather too many orthographic and grammatical errors for a monograph at this price, it can be recommended as the most convenient overview of this important phase of MT research.

M.T.Rosetta: **Compositional translation.** Dordrecht/Boston/London: Kluwer Academic Publishers, 1995. xviii,478 pp. ISBN: 0-7923-9462-3.

Behind the nom-de-plume of M.T.Rosetta, the researchers of the Philips Research Laboratories (Eindhoven, The Netherlands) have written the definitive account of one of the most innovative MT projects during the 1980s. The distinctive features of the Rosetta system are by now familiar to most of the MT community: in particular, the theory of compositionality applied to translation, the use of Montague semantics as the basis for an interlingua, and the reversibility of grammars. Besides this, the book is a substantial and important contribution in its own right to the theory of MT. It is not a collection of separately composed articles by members of the Rosetta group but the result of a genuine collaborative enterprise which reads as if written by a single hand. The first part contains chapters on the compositional definition of translation, M-grammars, the translation process, and the characteristic features of Rosetta. The second part provides further elaborations on morphology, dictionaries, syntactic rules, and controlled M-grammars. Part three comprises chapters on the linguistic aspects of the model. Part four looks at various translation problems: divergences between languages, temporal expressions, idioms, complex predicates, scope and negation. The final part considers in greater depth the formal aspects of M-grammars and aspects of software engineering. Clearly, this is an essential purchase for anyone interested in the linguistic foundations of MT and the contribution of this influential project.

PUBLICATIONS RECEIVED

Journals

Machine Translation Review: the periodical of the Natural Language Translation Specialist Group of the British Computer Society *1 (April 1995)*. Contents: pp.5-7: Group news and information. -- pp.8-9: Multilingual natural language processing (MNLP) project (David Wigg). -- pp.10-17: Machine Translation - Ten Years On: Cranfield conference report, 12-14 November 1994 (Derek Lewis). -- pp.18-19: CAT2 - a unification-based machine translation system (Ruslan Mitkov). -- pp.20-22: Practical aspects of the use of METAL at Siemens Nixdorf (Keith Roberts). -- pp.23-31: Linguistic resources on the Internet (Roger Harris). -- pp.32-36: Book reviews.-- pp.37-39: Conferences and workshops.

Computational Linguistics, *vol.20, no.4 (December 1994)* Contents: pp.507-534: A syntactic analysis method of long Japanese sentences based on the detection of conjunctive structures (Sadao Kurohashi and Makoto Nagao). -- pp.535-562: An algorithm for pronominal anaphora resolution (Shalom Lappin and Herbert J.Leass). -- pp.563-596: Word sense disambiguation using a second language monolingual corpus (Ido Dagan and Alon Itai). -- pp.597-634: Machine translation divergences: a formal description and proposed solution (Bonnie J.Dorr). -- pp.635-648: Training and scaling preference functions for disambiguation (Hiyan Alshawi and David Carter). -- pp.649-660: Storing logical form in a shared-packed forest (Mary P.Harper). -- pp.670-676: [Book review] Machine translation: a view from the lexicon, Bonnie J.Dorr (Daniel Radzinski).

Elsnews, *vol.4 no.1 (January 1995).* Contents include: Workshop on Machine Learning of Natural Language and Speech, report of ELSNET/MLnet workshop held in Amsterdam [2-3 December 1994] (Walter Daelemans); Need for European language resources infrastructure sparks negotiations (Joseph Mariani). *vol.4 no.2 (March 1995).* Contents include: MLAP projects in the area of transportation information systems (Roberto Billi); Multilingual automatic inquiry systems (Frank Seide); NLP research in Microsoft, Oz (Robert Dale)

INL Infoterm Newsletter *74 (December 1994).* Contents include (p.3) report on Workshop on "Language Engineering on the Information Superhighway" in Santorini (Greece), 26-30 September 1994 (Christian Galinski)

Language Industry Monitor *no.23 (September-October 1994).* Contents: pp.1-6: GSI-Erli. -- p.5: ALEP: a linguistic programming environment. -- pp.6-7: Putting Germany online. -- pp.7-8: Termbase with a twist [Linguistique Communication Informatique]. -- pp.8-9: A moving target [Eurolang Logos-Optimizer].

Language International, *vol.7 no.1 (February 1995).* Contents include: pp.29-33: Specialised dictionaries: expectations of users, practices of authors and publishers (Ad Hermans). -- pp.33-35: Euralex'94 (Ingrid Meyer). *vol.7 no.2 (April 1995).* Contents include: pp.11: CAT seminar in Slovenia (Sylvana Orel). -- pp.16-18: Communicating on Internet.-- pp.34-35: Major EC Language Engineering Programme.

Language Matters: news and views from ALPNET, *January 1995.*

LISA Forum Newsletter, *vol.3 no.4 (December 1994).* Contents include: pp.1-2: Translation memories, the impact on terminology ownership and business practice. -- pp.2-4: Vision of the translation market: an outline of the necessary reforms in Europe's language industry (Jaap van der Meer). -- pp.5-7: Business engineering and localisation framework (Judith Jones). -- pp.9-12: Genba wa tsuyoi. The strength is with the field, or Know your market! (Jan Pfefferkorn). -- p.14: Test suites for natural language processing (Lorna Balkan). -- p.15: The LISA Showcase [see this issue]. *vol.4 no.1 (March 1995).* Contents include: pp.1-2: Directions in the localisation industry (Rose Lockwood). -- pp.3-13: Highlights of conventional U.S. domestic software quality assurance (Emmanuel Uren). -- pp.13-14: Growing awareness of multilingualism - ITALICS sets the stage (Jaap van der Meer). -- pp.15-24: Document globalization: process and guidelines (Will Doherty).

Machine Translation, *vol.9 no.2 (1994/95).* Contents: pp.81-98: Augmenting formal semantic representation for NLP: the story of SMEARR (Victor Raskin, Salvatore Attardo, Donalee H.Attardo). -- pp.99-132: Multilingual dialogue-based MT for monolingual authors: the LIDIA project and a first mockup (Christian Boitet and Hervé Blanchon). -- pp.133-149: A simple and practical method for evaluating machine translation quality (Stephen Minnis).

Terminology, *vol.1 no.2 (1994).* Contents include: pp.253-275: On the empirical inadequacy of terminological concept theories: a case for prototype theory (Britta Zawada and Piet Swanepoel). -- pp.303-325: Terms and words: propositions for terminology (Isabel Desmet and Samy Boutayeb). -- pp.351-373: A model for the definition of concepts: rules for analytical definitions in terminological databases (Juan C.Sager and Marie-Claude L'Homme).

Terminology Standardization and Harmonization: newsletter of ISO/TC 37, *vol.6 no.4 (December 1994)*

The Translator, *vol.1 no.1 (1995).* Contents: pp.1-24: Translation, authorship, copyright (Lawrence Venuti). -- pp. 25-46: The moral dilemmas of court interpreting (Ruth Morris). -- pp.47-64: Observations on anomalous stress in interpreting (Sarah Williams). -- pp.65-86: A descriptive framework for compensation (Keith Harvey). -- pp.87-92: The dawn of a modern theory of translation (Juan Sager).

Books

Gawronska, Barbara: **An MT oriented model of aspect and article semantics.** Lund: Lund University Press, 1993. (Travaux de l'Institut Linguistique de Lund 28) 246pp. ISBN: 91-7966-237-4.

Zampolli, Antonio; Calzolari, Nicoletta; Palmer, Martha (eds.) **Current issues in computational linguistics: in honour of Don Walker.** Pisa: Giardini, Dordrecht: Kluwer, 1994. xxv, 595pp. (Linguistica Computazionale, vol. 9/10). ISBN: 0-7923-2998-8. Contents include: On the proper place of semantics in machine translation (M.King). -- Construction-based MT lexicons (L.Levin, S.Nirenburg). -- Stone soup and the French room (Y.Wilks).

Kitano, Hiroaki: **Speech-to-speech translation: a massively parallel memory-based approach**. Boston/Dordrecht/London: Kluwer Academic Publishers, 1994. xvii,192pp. ISBN: 0-7923-9425-9.

Weisweber, Wilhelm: **Termersetzung als Basis für eine einheitliche Architektur in der maschinellen Sprachübersetzung**. Das experimentelle MÜ-System des Berliner Projekts der EUROTRA-D-Begleitforschung (KIT-FAST). Tübingen: Niemeyer, 1994. (Sprache und Information, Bd.28) xviii,262pp. ISBN: 3-484-31928-3

Sigurd, Bengt (ed.): **Computerized grammars for analysis and machine translation.** Lund: Lund University Press, 1994. (Travaux de l'Institut de Linguistique de Lund 29) 148pp. ISBN: 91-7966-304-4.

Rosetta, M.T.: **Compositional translation.** Dordrecht/Boston/London: Kluwer Academic Publishers, 1995. xviii,478 pp. ISBN: 0-7923-9462-3.

Sinclair, John; Hoelter, Martin; Peters, Carol (eds.): **The languages of definition: the formalization of dictionary definitions for natural language processing**. Luxembourg: European Commission, 1995. (Studies in Machine Translation and Natural Language Processing, Vol.7) 209pp. ISSN: 1017-6568

Conference proceedings

CSNLP 1994: 3rd Conference on the Cognitive Science of Natural Language Processing, Dublin City University, 6-8 July 1994. Editor: A.I.C.Monaghan. Dublin: Natural Language Group, Dublin City University.

Second Annual Workshop on Very Large Corpora (WVLC2). Program and proceedings, Thursday, 4 August 1994, Kyoto International Community House, Kyoto, Japan. 159pp.

KONVENS '94: Verarbeitung natürlicher Sparche. Tagungsband. Hrsg. H.Trost. Wien: Österreichische Gesellschaft für Artificial Intelligence, 1994. ix,442pp. (Informatik Xpress 6)

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 in July 1993, which asks all members to send copies of all their publications within one year of publication.