MT News International

Newsletter of the International Association for Machine Translation

ISSN 0965-5476 Issue no.5, May 1993

Editor-in-Chief:
John Hutchins, The Library, University of East Anglia, Norwich NR4 7TJ, United Kingdom
Fax: +44 (603) 259490; Email: L101@uea.ac.uk
Regional editors:
AMTA: Joseph E.Pentheroudakis, Microsoft Corporation, Redmond, WA 98052, USA
Fax: +1 (206) 936-7329; Email: josephp@microsoft.com
EAMT: Tom C.Gerhardt, CRPCU/CRETA, 13 rue de Bragance, L-1255 Luxembourg.
Fax: +352 (44) 73 52; Email: tom@crpculu.lu
AAMT: Professor Hirosato Nomura, Kyushu Institute of Technology, Iizuka, 820 Japan.
Fax: +81 (948) 29-7601; Email: nomura@dumbo.ai.kyutech.ac.jp

Published in the Americas, with the generous assistance of Microsoft Corporation

CONFERENCE REPORTS

Building Lexicons for Machine Translation

Clare R. Voss

This year in its annual spring ritual at Stanford University, the American Association for Artificial Intelligence (AAAI) included among its spring symposia one entitled "Building Lexicons for Machine Translation." The symposium, chaired by Bonnie Dorr and organized with the assistance of Michael Brent, Sergei Nirenburg, Elaine Rich and Patrick Saint-Dizier, provided a refreshing mix of panel discussions, invited talks, paper presentations, and outdoor refreshment breaks. James Hendler chaired the symposium series this year and will co-chair it with Lynn Stein next year.

Possibly the most entertaining moments of the meeting came when Sergei Nirenburg gave his 10-minute overview of the symposium during the plenary session. In his words, the burning issues being raised by the MT practitioners were:

(1) content: how little can we get away with?
(2) form: how dare you say that your formalism is better than mine?
(3) provenance: wouldn't it be nice if somebody else did it for us?
(4) clients: must we really be bothered by all these messy phenomena?

No doubt one reason Nirenburg raised the content issue was that he noted several presentations where the lexical representations were encoded minimally. At this symposium we heard from several groups working on automatic acquisition (i.e., information extraction and analysis) from corpora and dictionaries. Researchers such as Chinatsu Aone and Doug McKee, Megumi Kameyama, Stanley Peters, and Hinrich Schuetze did not invoke a cry for minimalism on theoretical grounds, but rather implicitly on practical grounds. Finding that they need to provide broad coverage in new domains quickly, they have opted for a minimal initial approach that they can then gradually fine-tune over time. Patrick Saint-Dizier addressed the processing side of this issue and described constraints that produce minimal lexical information passing in his system. At the other end of the spectrum were efforts directed at representing fine-grained lexical distinctions. Chrysanne DiMarco, Graeme Hirst and Manfred Stede have been exploring ways of formalizing the language of dictionary usage notes for capturing meanings of near-synonyms.

The topic of preferred formalisms, the second of Nirenburg's issues, arose following Ann Copestake and Antonio Sanfilippo's presentation of 'tlinks' (translation links). This mechanism is used to link ‘typed feature structures’, the formal structures used for word entries as well as other linguistic units, such as phrases. Christian Rohrer asked for help in finding arguments in support of the typed formalism over general feature systems without typing. Aside from enforcing well-formedness constraints, Copestake pointed out that typing facilitates identifying errors and locating them. In contrast to this work where the same formal structure is used for syntactic and semantic information within a single lexical entry, Modiano described the MULTILEX open structure approach to lexical entries. Additional formalisms were presented in two papers, one by Bianka Buschbeck-Wolf and another by Bonnie Dorr and Clare Voss, both of which focused on the translation of spatial expressions. In the first case, the formal language described was a set-theoretic notation developed for translating between prepositions in identical syntactic structures (a noun phrase containing a modifying prepositional phrase). By contrast Dorr and Voss presented a formalism for ‘spatial predicates’, predicate-argument structures for spatial relations at a lexical-semantic, or interlingual level of representation. Another focus for formalization efforts were the metrics used in decision algorithms to select the ‘best’ translation from the class of output sentences. James Barnett and Elaine Rich presented their work developing measures of ‘semantic accuracy’ and ‘semantic naturalness.’

Nirenburg's third issue is a curious variation on the classic divide-and-conquer strategy, where he suggests that some MT researchers are splitting up the work in order to avoid conquering the tough problems. Perhaps it was this observation that originally led to the work presented by Lori Levin and Sergei Nirenburg at the symposium. In their analysis, they describe a continuum of translation cases that extends from those that can be handled by general, productive, predictable principles in MT systems, out to the non-productive, idiosyncratic cases that must each be handled in some specific way.

Perhaps to inspire us to look beyond the usual places and not to give up on those unruly cases, Dorr invited two linguists not widely known in the MT community, to speak about their work in lexical semantics. Beth Levin, the first to speak, presented some of her analyses of intransitive verbs and their semantic groupings. Levin ended her talk by suggesting that this phenomenon of ‘regular polysemy' might also extend in systematic ways cross-linguistically, thereby guiding the construction of lexicons.

The second speaker, Alan Cruse, presented his theory of polysemy. After many amusing sentences illustrating ways to differentiate interpretations of polysemous noun phrases, Cruse documented inconsistencies and inadequacies in James Pustejovsky's Qualia theory. This also seeks to account for polysemy and is more widely known in the computational linguistics community. Finally he speculated on ways that Qualia theory and prototype theory could be brought together with his work.

The panels organized by Scott Bennett and Sergei Nirenburg, as well as the invited talk by Makoto Nagao, provided a forum for discussion about Nirenburg's fourth issue, namely the struggles underway and the resources currently available to MT system developers and users in the commercial world and at academic institutions. Bennett and several panellists pointed out that nobody had yet addressed two types of problems related to those messy client concerns. One was the need to develop MT systems that are robust enough to cope with newly-coined terms and phrases (eg., noun-noun phrases in government documents) and ungrammatical input from technical texts. The second was the need to let users do some of the lexicon customizing themselves.

The wide range of motivations for extracting information from on-line resources was quite striking in one morning session. Virginia Teller and Eleanor Olds Batchelder presented an algorithm to segment unrestricted hiragana text into words for a study of Japanese morphology. Stephen Helmreich, Louise Guthrie, and Yorick Wilks described the Pangloss project's use of machine-readable dictionaries to build extensive lexicons with phrases and idioms as entries that will eventually merge with corpus-derived lexicons. Dirk Heylen, Kerry Maxwell, and Susan Armstrong-Warwick discussed their linguistic investigation into the semantics of collocations that appear implicitly in a learner's dictionary and in bilingual dictionaries.

The symposium ended on an unexpected note. Susann Luperfoy, who had just been at an ARPA workshop on lexicon issues, told the audience that the COMLEX project leaders want input from the MT community on defining what belongs in the dictionary they are building.

For more details about the symposium, readers may order Technical Report SS-93-02 from the AAAI at 445 Burgess Drive, Menlo Park, CA 94025, USA. Email: sss@aaai.org. [Details given in the 'Publications Received' section.]

Machine Translation and the Lexicon

Petra Steffens and John Hutchins

The third international workshop of the European Association for Machine Translation (EAMT) took place from the 26th to 28th April 1993. Its topic was "Machine Translation and the Lexicon." It was organised by Petra Steffens (IBM Information Systems, Germany) in cooperation with Ulrike Schwall (Siemens Nixdorf Information Systems), and was hosted by the IBM Scientific Centre in Heidelberg. The workshop was divided into three sessions: Types and representation of lexical information; Acquisition, organisation and maintenance of lexical information; Terminological databases.

It is hoped to included a detailed report of the workshop in the next issue of MTNI. In the meantime, the following list of speakers, titles and brief abstracts give some indication of the coverage. Publication of the proceedings of the workshop is currently in preparation.

Session I (April 26) -- Nicoletta Calzolari (Univ.Pisa): 'European efforts towards standardizing language resources'. Nicholas Ostler (Linguacubun, London): 'Perception vocabulary in five languages - towards an analysis using frame elements' [on the DELIS project involving Danish, Dutch, English, French and Italian, within framework of Fillmore's frame semantics.] Ulrich Heid (Univ. Stuttgart): 'Using typed feature structures for the representation of syntactic lexical semantic information in a contrastive lexicon' [also on the DELIS project, with details of the unification-based syntactic description model, and concentrating on divergences and mismatches in the lexical fields of movement and perception.] Christa Hauenschild (Univ. Hildesheim): 'The lexicon and the general design of MT systems'. Hans-Ulrich Krieger (DFKI Saarbrücken): 'Derivation without lexical rules' [on value of typed-feature formalisms (of HPSG kind) for representing the effects of lexical rules, for integrating allomorphology and morphotactics, and for simplifying logical form.] Peter Hellwig (Univ. Heidelberg): 'Towards re-usable lexical resources'.

Session II (April 27) -- Folker Caroli (IAI Saarbrücken): 'Types of lexical co-occurrences - descriptive parameters' [on treatment of multi-word units, idioms, collocations, support verb constructions, and their representation in large-scale reusable lexical resources.] Brigitte Bläser (IBM Heidelberg): 'Lexicon and terminology management in the TransLexis project' [in the environment of Translation Manager/2 and the integration within it of the prototype MT system LMT for English-German and German-English translation.] Ulrike Schwall (SNI, München) and Angelika Storrer (Univ. Tübingen): 'Description and acquisition of multiword lexemes'. R.Lee

Humphreys (SITE-Eurolang, Maisons Alfort): 'Lexical models for MT and MAHT'. Renate Mayer (FAO Stuttgart): 'Navigation in terminological knowledge bases'. Jean Véronis (CNRS Marseille): 'From dictionaries to knowledge bases (and back?)' [combining information from multiple sources and exploiting dictionary structures by non-symbolic methods.] Michael Anobile (LISA Geneva): 'LISA, its role in localization and multilingual documentation'.

Session III (April 28) -- Wilhelm Daelemans (Univ. Tilburg): 'A generic architecture for data-oriented lexical acquisition' [involving case-based reasoning (instance-based learning) and information entropy.] Patrick Burkhard (Union Bank of Switzerland): 'Terminology at Union Bank of Switzerland' [development of TerMS, with one database for each language and multilingual interrogation.] Christian Galinski (TermNet Vienna): 'Machine translation and terminological standards'. Khurshid Ahmad (Univ. Surrey): 'Management of large linguistic resources for knowledge processing' [on data models appropriate for large volume lexicons, from experience with the Translator's Workbench projects and MULTILEX.] Vangelis Karkaletsis (NCSR 'Demikritos', Univ. Athens): 'Software localization and the GLOSSASOFT project' [an LRE project.] Paul Procter (Cambridge Univ. Press): 'Canbridge Language Survey: the development of a non-language specific semantic coding system using multiple inheritance' [a model of the mental lexicon, including non-scientific and non-taxonomic features, coexistence of different classifications and multi-dimensional semantic hierarchies.] Nancy Ide (CNRS Marseille): 'Database models for dictionaries' [object-oriented database system.] Katharina Koch and Daniel Grasmick (SAP Walldorf): ' How to avoid inconsistencies when using a traditional and a machine translation TDB - as attempted by SAP' [use of METAL at German software company.]

Technology and Language in Europe 2000

John Hutchins

On 15th January, Logica Cambridge and UnivEd Technology Edinburgh organized a one-day symposium on behalf of DG XIII-E of the Commission of the European Communities. Under the title *Technology & Language in Europe 2000 - the UK perspective*, the meeting was intended to promote the Commission's most recent initiatives in the area of Linguistic Research and Engineering (LRE). Invitations were sent to senior figures in commerce, industry, government, and research to participate in the ‘briefing session’, chaired by Brian Oakley.

The first half of the morning was devoted to presentations of CEC and governmental views. André Danzin spoke as author of the CEC report `Towards a linguistic infrastructure for Europe' on the need for supra-national support of language technologies in Europe. Frans de Bruïne outlined the CEC strategy for stimulating and promoting the application of promising technologies for the improvement of European competitiveness. Progress in language technologies was also important for democracy in the Community, ensuring that all could participate, and for supporting minority languages. There was real danger of US dominance in this area. Hervé Gallaire (Rank Xerox) spoke of the need to focus on a limited number of technologies with most application potential. Gerry Gavigan described the UK Department of Trade and Industry's involvement in research on speech and language technologies and the encouragement of industrial collaboration. Jan Roukens outlined the ‘LE 2000’ programme to support research in multilingual computer-assisted services, particularly in telecommunications, and to promote the formulation of the necessary standards. Rose Lockwood (Ovum Ltd) presented the findings of a survey of the United Kingdom market for language technologies: natural language access to databases of information and communication networks, speech and voice processors, and translation tools and systems. The UK lagged behind the rest of Europe regarding language ability of company staffs, some 40% reporting problems from inadequate language skills. She was followed by Geoffrey Kingscott and Veronica Lawson, who described the development of MT and computer-assisted translation systems and the attitudes of the translation profession to these developments; in particular, training courses should be more oriented to language technologies.

In the afternoon, the symposium turned to some specific applications and research projects. Doug Embleton of ICI gave a businessman's view of an integrated strategy involving language engineering products. Peter Pym described the successful use of a controlled English for documents to be submitted to an automatic translation system at Perkins Engines; and Gerhard Freibott described the integration of translation technologies at Krupp Industries. Sir Roger Elliott gave a view from Oxford University Press of developments in electronic publishing; and Bruce Bond reported on the activities of British Telecom in areas relevant to language technology. In the final session, Steve Pulman (Cambridge University) described the Combined Language and Reasoning Engine (CLARE), a powerful research tool with various potential applications in the area; Jeremy Peckham (Logica) reported on current research in speech technology; and John Laver (University of Edinburgh) speculated on future markets for language technologies in multilingual Europe. Equality of access to information is a basic social right which should not be hampered by language barriers. Multilingualism needs to be encouraged and not discouraged by technological and market-led convergence on `major' languages. The day ended with a panel chaired by Nick Ostler for symposium participants who were not users of any language technologies to say whether what they had heard was likely to change their views. The panellists represented Amnesty International, British Rail Passenger Service and British Aerospace. The overall message was that they needed more convincing.

Proceedings are available (priced at £20) from Ian Rae, Logica Cambridge Ltd., 68 Newman Street, London W1A 4SE.

Similar events on behalf of the Commission of the European Communities took place on 16-17 February 1993 at the Abbaye des Prémontrés, *Pont à Mousson*, France, organised by the Observatoire Français des Industries de la Langue, the French Eureka chairmanship and the regional council of Lorraine; and on 10-11 May 1993 (Forum des Industries de la Langue), *Brussels*, organised by the Observatoire Wallon des Industries de la Langue and Centrum voor Computerlinguïstiek. It is hoped to report on these events in the next issue of MTNI.

EACL-93, Utrecht, 21-23 April 1993

John Hutchins

The sixth conference of the European Chapter of the Association for Computational Linguistics got off to a lively start with a warning call from Ken Church (AT&T Bell) to the MT research community. If it does not change its objectives it will lose its support. MT research is far too theoretical (and in this he included statistics-based MT) and not designed for serious practical use in the near or even distant future. What it should aim for is the development of systems which can help human translators in time-consuming tasks. It should actively solicit the assistance of translators, watch what they do, use their experience, get them to evaluate and test systems. If they can be given help now, however imperfectly, then they will back MT research for developing future aids. There are some aids already available or imminent: termbanks, terminology managers, storage of translations for reuse or consultation, translation workstations, etc. (Church mentioned the work at AT&T on the TermWorks workstation and on automatic alignment of original and translated text segments.) But even the translation workstations, linguistically modest and crude as they are, may be on the wrong track. More than half their time, translators are away from their desks checking sources and terminology, consulting reference books, etc. The translation process itself represents no more than 30% of the total time and cost of producing translated texts. Productivity gains have come more from improved text processing, inputting, formatting and printing than from any automated translation aids. We need the active support of translators to make progress, to justify what really needs to be done, and not to do what translators can do much better.

While many of the papers at EACL93 were of interest to MT research, there were fewer this year directly concerned with translation itself. Ron Kaplan and Jürgen Wedekind spoke about problems in LFG relevant to translation (specifically the head-switching problem illustrated by the structural misalignment of, e.g., *Le bébé vient de tomber* and *The baby just fell.*) Josef van Genabith gave a paper (by Doug Arnold et al.) on "Experiments in reusability of grammatical resources" by the development of methods for migration from HPSG, ETS (Eurotra) and LFG formalisms into the general-purpose ALEP formalism. Donna M.Gates and Peter Shell described a system at Carnegie-Mellon University for the semi-automatic acquisition and maintenance of lexical and semantic knowledge. The COOL system (Creator of Ontologies and Lexicons) was developed for ESTRATO, a project for translating from Spanish into English using various modules of the CMU KANT system and the CMU TWS workstation. In a student session, Hadar Shemtov (Stanford University) described his research on text alignment as a tool for translating revised documents - an example of the practical work advocated by Ken Church. In a demonstration session, Shinichi Doi reported on the NEC method for preliminary segmentation of long sentences (in Japanese newspapers) by a combination of pattern matching and keyword identification. In another demonstration, Eric Wehrli (University of Geneva) showed his prototype PC-based system ITS-2 featuring interactive analysis of French (for monolingual users), a Government-Binding based parser, simple bilingual transfer, and elementary transformations for generating English output. A demonstration of the Carnegie-Mellon Pangloss system could unfortunately not take place.

The proceedings of the conference are available from Donald E.Walker (ACL), Bellcore, 445 South Street MRE 2A379, Morristown, NJ 07960, USA.

ASSOCIATION NEWS

EAMT General Assembly, Utrecht, 22 April 1993

Ian Johnson

Minutes taken at the General Assembly of the EAMT held at 18:00 on *Thursday 22 April 1993* at CSB, Utrecht, Netherlands.

Present:Maghi King, Doris Albisser, Ian Johnson, John Hutchins, and 23 other members of the General Assembly of the EAMT.

The agenda was drawn up and accepted.

1. **Reports from Provisional Committee**

1.1 *Presidents Report*. Maghi King reported that the EAMT had passed its target of 100 members in 1992, and already organised a number of workshops. The aim now is to build on this achievement and develop the publicity and PR side of the EAMT's activities. A major event on the horizon is the participation in organisation of the MT Summit V to be held in Europe in 1995.

1.2 *Secretary's Report*. Ian Johnson reported the following details about the membership of the EAMT at the end of 1992 and the current status:

	1992	1993 (at 21/4/93)
Individual	82	64
Non-Profit Making Institutions	7	3
Company/Profit-Making Institutions	14	10
TOTAL	103	77

Fees are currently as follows:

Individual	SFr 35
Non-Profit Making Institutions	SFr 175
Company/Profit-Making Institutions	SFr 350

1.3 *Workshops Report*

1.3.1 A report on the EAMT Workshops was prepared by Ulrike Schwall and read out by Ian Johnson, as she was unable to attend the meeting.

1.3.2 A Workshop on Machine Translation and Translation Theory was held in Nantes on 22 July 1992. It was organised by Christa Hauenschild (University of Hildesheim) and Ulrike Schwall (Siemens Nixdorf). About 40 participants attended.

1.3.3 A Workshop on Computer Integrated Translation and the User was held in Saarbrücken on 29-30 October 1992. It was organised by Tom Gerhardt.

1.3.4 A Workshop on Machine Translation and the Lexicon will be held in Heidelberg from 26-28 April 1993 organised by Petra Steffens (IBM) and Ulrike Schwall (Siemens Nixdorf).

1.3.5 The EAMT is lending its patronage to the Bulgarian School on Computational Linguistics which will be held in Bulgaria on 5-11 September 1993, organised by Ruslan Mitkov.

1.3.6 Further Workshops are being planned for 1994.

1.4 *Treasurer's Report*. Doris Albisser presented the final EAMT accounts for the last financial year dated 31 December 1992. Total income from membership fees and bank interest: SFr 8767.80. Total expenses, consisting mainly of newsletter costs and a 10% contribution to the IAMT: SFr 3708.60. This gave a net profit for 1992 of SFr 5059.20. Doris also gave information on the current status of the accounts for 1993 until the end of March, but requested that these not be written up in the minutes. Finally, she appealed for individuals to cover all bank charges associated with the payment of membership fees, because otherwise the EAMT accounts would lose too much money.

2. **Newsletter Report**

2.1 John Hutchins gave a summary of the status of the MT News International which appears three times a year. A recent innovation is that the Newsletter is available to non-members of the EAMT at SFr 70.

2.2 The IAMT is hoping to produce a directory of MT systems. As John Hutchins is responsible for editing this list, he requested that people send him any details they have about products available (brief technical details, languages in the system, number of words in the dictionaries, hardware required, software compatibility, and cost, etc.).

2.3 As far as the content of the Newsletter itself is concerned, he appealed for input from manufacturers and users. From the January 1994 issue, he is hoping to include advertisements placed by developers, etc. This is being organised by Bill Fry in the USA. He ended with a general request for more information about anything of relevance to the MT world.

3. **Adoption of the Statutes**

3.1 The EAMT Statutes were formally adopted by unanimous vote of the General Assembly.

4. **Elections to the Committee**

4.1 Maghi King announced the list of nominations to the Committee. Five valid nominations were received.

4.2 The General Assembly elected the following people to the Executive Committee: Maghi King (President), Doris Albisser, Ian Johnson, Veronica Lawson, Ulrike Schwall

4.3 Maghi King announced that the Committee had been elected for a period of three years.

5. **Membership Fees**

5.1 It was decided to keep the current individual membership at the relatively low rate of SFr 35. Doris Albisser reminded everybody that payment can be made in local currency, but that all bank charges incurred should be paid by the member. Individual members should always give their home address for mailing purposes.

5.2 Ruslan Mitkov from Bulgaria suggested that the institutional rate for ex-Eastern bloc countries should be lowered. The General Assembly voted in favour of giving the Committee the authority to change the membership payment system accordingly, to take into account hardship cases. A section would be added to the membership form allowing people to apply for special discounts by writing a letter explaining their situation.

5.3 A proposal to set up an International Fund, allowing better-off members to pay an amount in addition to their membership fee to go towards a fund to support the less well-off members. This suggestion was accepted by the Assembly, but it was decided that the EAMT should put the proposal to the next meeting of the IAMT in Japan in July before setting up such a fund.

6. **Site of MT Summit**

6.1 Loll Rolling (Head of Division, Linguistic Engineering at the European Commission) presented the formal proposal from the EC that MT Summit V should be held in Luxemburg in 1995, with EC as sponsor. He has been actioned to do a feasibility study and cost structure and to produce an action plan. Serge Perschke will also be kept aware of progress, as Loll Rolling is retiring in 1994.

6.2 The suggested location of the Conference is Luxemburg. The suggested dates are either 27-29 June 1995 or 9-11 July 1995. Loll Rolling mentioned that there seemed to be a slight preference among university professors for the earlier date, but thought industry might prefer the later date. It will be necessary to find out the dates of the Theoretical and Methodological Issues in MT (TMI-95), to ensure there is no clash.

6.3 Loll Rolling has already carried out some preliminary exploration of conference locations and facilities, transportation and accommodation possibilities in and around Luxemburg, together with a comparative cost evaluation. A cost estimate has already been received from F.I.L. (Société des Foires

Internationales de Luxembourg) for the rent of conference and exhibition rooms, including coffee breaks and lunch. Another cost estimate from the city of Luxemburg concerning the E.P. hemicycle should arrive shortly. The mayor of Luxemburg has written a letter offering support for the event (which is expected to take the form of a reception paid for by the city. An estimated participation fee of about 250 ECU for the average participant will not completely cover costs, so a 60000 ECU subsidy from the EC has already been agreed (more may be available if necessary, but this will have to be separately negotiated).

6.4 The other principal actions which have to be taken in the course of this year are the nomination of a steering committee consisting of European Commission and EAMT representatives, nomination of a programme committee composed of MT experts, and negotiation with ITC, a Luxemburg-based company who could handle the conference management. The first meeting of the steering committee is scheduled for September 1993.

6.5 A vote was taken and passed unanimously in favour of accepting the EC proposal to hold MT Summit V in Luxemburg in 1995.

6.6 It was agreed that the proposal would be discussed in more detail by Loll Rolling and members of the EAMT Executive Committee at a meeting to be held the following morning (23 April 1992 at 10:30 a.m.) at the CSB in Utrecht.

7. **Matters Arising**

7.1 The proposal for the MT Summit was further discussed by the General Assembly. Some people felt that the average proposed participation fee of 250 ECU sounded a bit high and suggested that the academic fee should be reduced and the industrial fee increased. Harry Somers pointed out that English↔Japanese and French↔Japanese interpreters would probably be more useful than just the English↔French interpreters that had been mentioned in the proposal.

8. **Any Other Business**

8.1 A request was received from Dan Cristov from Rumania for the EAMT to be Honorary sponsors of the Rumanian Summer School later this year. Loll Rolling pointed out that the EC has money to support this sort of activity. In the course of the discussion it emerged that the content of the Summer School was AI and Computational Linguistics, but that there would be no specifically MT content. It was argued that the EAMT should focus primarily on MT-related activities. The decision was therefore to recommend that the EACL lend their support to this Summer School and that the EAMT would be prepared to support a future Summer School in Rumania if it was concerned with MT.

8.2 Various places were suggested for the next General Assembly (Aslib, COLING). No final decision was reached. Harry Somers suggested that we should organise an MT conference next year, since there are currently no conferences on MT scheduled for 1994. This suggestion was well-received and it was agreed that the Committee should pursue it further. One possibility which was mentioned was that there was a plan to internationalise the German MT User Group and that this could in some way be linked to such a Conference.

Ian Johnson (Secretary, Executive Committee of the EAMT)
30 April 1993

PEOPLE ON THE MOVE...

As many will already know, *Yorick Wilks* is moving back to the United Kingdom in July. Having been director of the Computing Research Laboratory at the New Mexico State University since 1985, involved in many areas of important research well known within the AI and CL communities, he is taking up a professorship in the Department of Computer Science at Sheffield University.

Job van Zuijlen has left BSO in Utrecht to be a Senior NLP researcher at the SRA Corporation in Arlington, Virginia, USA. He will be a member of its Intelligent Information Systems division, involved in research on knowledge acquisition using corpora and statistical techniques and thus perhaps continuing some of the work undertaken for the DLT project. His email address is: zuijlenj@sra.com.

SYSTEMS and PROJECTS

News from LOGOS

Logos Corporation announce that it has added Italian to the targets available on its English multi-target machine translation system (Release 7.5; other targets are French, German and Spanish). The Italian target was co-developed with Thamus, the Italian member of Eurolang in located in Salerno. The English multi-target system runs under UNIX on a SPARCstation and is fully integrated with the major electronic publishing and word-processing systems.

Logos Corporation has an immediate opening for a development linguist with native French and near-native German. This person would be trained to take on major responsibility for French target in the German multi-target system. Training would take place at the company's U.S. technology center in New Jersey with eventual assignment in Germany. Logos Corporation is also looking for entry-level linguists with native Spanish, French, Italian and German.

Logos Computer Integrated Translation GmbH (formerly Logos Computer Systems Deutschland, GmbH) has relocated to Eschborn/Taunus, right outside of Frankfurt. A subsidiary lexicographical operation has just been opened in Berlin, aimed at the construction of specialized lexicons for the Logos MT system. Logos has also entered into a close working relationship with Prof. Haller and members of his institute (IAI) at the University of Saarbrücken. Thus far, activities have focused on evaluation of the Logos German multi-target system (with English/French/Italian targets). The collaboration between Logos and IAI is expected to expand in the near future to include joint development activities. Until the end of last year, Haller's IAI was the centre for Eurotra-D activities in Germany.

The address of Logos Corporation in the U.S. is: Logos Corporation, Suite 214, 111 Howard Boulevard, Mt. Arlington, New Jersey 07856. Tel: (201) 398-8710, fax: (201) 398-6102.

LMT demonstrated at CeBIT 93

[Extract from *Language Industry Monitor 14*]

CeBIT 93's most pleasant surprise was to discover Hubert Lehmann of IBM Deutschland's Heidelberg labs demonstrating an English-German prototype of IBM's Logic-Based Machine Translation (LMT) system. Lehmann was sequestered away in one of IBM's busy offices on the roof of Halle 1, the CeBIT-dedicated building at the Hannover Messe in which a number of the large exhibitors have erected permanent edifices. Lehmann was using IBM's Translation Manager/2 package as a front end to LMT, which was running on a mainframe in Heidelberg. Lehmann stressed that LMT services could be integrated within TM/2, ideally sharing the translation memory data and user dictionaries of that OS/2 package. Is LMT the first sign of IBM's long-awaited entry into the MT market?

LogoVista E to J

In January, Language Engineering Corporation (LEC) announced the introduction of LogoVista E to J, the first of the LogoVista family of translation support systems. LogoVista E to J is a general-purpose system which can translate a wide variety of English text into Japanese.

The system features a 100,000-entry main dictionary which can be supplemented by any of 19 optional technical dictionaries. The technical dictionaries, which cover technical terms from many fields, contain a total of over 350,000 entries. LEC developed the core technology, while LEC's Japanese partners developed the technical dictionaries. LEC markets and sells LogoVista E to J in the United States, while LogoVista Corporation, a joint venture formed by LEC, CATENA

Corporation (ComputerLand Japan), and Risousha Incorporated, markets and sells the system in Japan.

LogoVista E to J is available for the Macintosh and for Sun Microsystems, Hewlett-Packard, and Sony Microsystems UNIX workstations.

A Japanese Windows version will be available in July, 1993. A Japanese to English version of the software, LogoVista J to E, is under development.

For more information, please contact Language Engineering Corporation, 385 Concord Avenue, Belmont, MA 02178, tel: +1 (617) 489-4000, fax: +1 (617) 489-3850.

PC-Translator

A new version, PC-TRANSLATOR 3.4, was introduced by Linguistic Products of The Woodlands, Texas, last fall at the MT Showcase in San Diego. Its main improvement, as reported in MTNI#4, is the ability to accept multiple Wildcard Phrases. In order to take full advantage of this feature, Linguistic Products has refined its dictionary coding and is currently engaged in a complete overhaul of all their language pairs.

The company reports that, by using a computerized technique, the overhaul of three language pairs, namely, English-French, English-Spanish, and English-Swedish, has already been completed. A new English-Portuguese version was also introduced last month, raising the total of PC-TRANSLATOR language pairs available to twelve.

PC-TRANSLATOR's ability to maintain input format also has a new application in the generation of multilingual forms originally created in Microsoft Word or WordPerfect. According to the company, in this application PC-TRANSLATOR can be made to reproduce existing translations or create new ones, thus reducing significantly the cost incurred by typesetting duplication and facilitating revisions and updates.

The company also reports that two magazines in Europe and one in South America are preparing articles comparing PC-TRANSLATOR to other MT systems.

For more information, please contact Linguistic Products, P.O. Box 8263, The Woodlands, Texas 77387, telephone (713) 363-9154, fax: (713) 298-1911.

News from the Confederation of Independent States

Evgenii Lovtskii reports that research on MT continues under severe practical and financial adversities in the former Soviet Union. Information is difficult to gather, but here are some of the projects active in some way at the present time.

At the Centre for Translation in Moscow (formerly VCP: All-Union Centre for Translation), Boris Tikhomirov and his group are completing an English-Russian one-directional system for an IBM PC AT machine. It is a replica of the AMPAR system which was in service for many years in several institutions on Soviet-made mainframes. It is to be demonstrated at a software exhibition in mid-April.

Zoya Shalayapina continues development work on her Japanese-Russian system at the Institute of Oriental Studies. The system makes extensive use of semantics, including a special semantic representation. Unfortunately, organisational difficulties are endangering the whole project.

The once numerous and influential group 'Statistika Rechi' (Speech Statistics) of Leningrad (now St Petersburg), headed by Raimund Piotrowski, has decreased in numbers and has finally split into two groups, each of which has come forward with an MT system. Both are commercially available and are said to be quite good.

The PARS system by Misha Blekhman (Kharkov) - a two-way English-Russian MT system - is now available for purchase in St Petersburg, Kharkov and Moscow. Apparently ten have been sold. Recently he has developed a Russian-Ukrainian system. Its value in the now bilingual state of Ukraine is indicated by more than 70 installations of the system.

It is known that other places where research groups are still active include: the University of Tver' (formerly Kalinin), Kiev under G.Miram, and Kishinev (now capital of the Republic of Moldavia) where a group under G.Nikolayev developed an English-Russian system, which is now being transported to a PC.

Lovtskii hopes to be able to report further on these activities in a future issue of MTNI.

PANGLOSS developments

[Edited extracts from an interview with Eduard Hovy published in *Language Industry Monitor 14*]

One of the most ambitious interlingua MT systems to date is the PANGLOSS project, a collaboration between three American research groups with funding coming from ARPA. Initially the researchers are working on a Spanish-English system, it will be followed eventually by a Japanese-English one. For the analysis stage of PANGLOSS, New Mexico State University's Computing Research Laboratory (CRL) is contributing its ULTRA parser, which will work in tandem with a Spanish lexicon tagged with the Longman's LDOCE sense keys; this will be the initial bilingual lexicon for the system. At Carnegie-Mellon University in Pittsburg, the Centre for Machine Translation is developing the interlingua for PANGLOSS. Here, the output from CRL's parser will be mapped to a complex scheme of case frames, in which such entries as 'agent', 'theme' and 'experiencer' are identified. These are mapped to the database of 'concepts' - individuals, events, or states, which in turn have their own attributes and are linked causally, temporally, or spatially to other concepts. There are slots in these concepts where discourse and other kinds of pragmatic information can be tracked... Much os this is based on AI research from the past decade; the CMU MT Centre worked for many years at applying AI techniques to MT. The resulting Knowledge-Based Machine Translation (KBMT) technology is also at the heart of the ambitious fifteen-year project the Centre is now undertaking with the Carnegie Group at the behest of Caterpillar.

Working in close collaboration with the Carnegie Mellon and the New Mexico groups, the Information Sciences Institute (ISI) at the University of Southern California is developing the module to generate the English output in PANGLOSS. ISI is regarded as the bastion of text generation, primarily on account of PENMAN, its text generation system. PENMAN has been very generously funded over the past decade and it now circulates freely within the research community. As Eduard Hovy (leader of the ISI group) explains, because generation is more straightforward than parsing, PENMAN is reasonably accurate. When parser encounter something that is not defined, they generally fail; generators, however, can fail gracefully by incorporating the unknown data literally and proceeding onward. A system like PANGLOSS will display the problematic data interweaved through the text. This is where the user comes in.

The PANGLOSS prototype is intended to be used interactively, with users resolving ambiguities and choosing formulations in the source language at a workstation being developed at CMU. The Caterpillar system is being designed for Caterpillar's Simplified English. Will PANGLOSS also be able to take advantage of the constraints imposed by some form of restricted input? "No," says Hovy, "that's the distinction between *dissemination* and *assimilation* in MT. Caterpillar wants to disseminate information and can afford to impose input restrictions. Our task is to gather information. We have to be prepared for any kind of input, albeit within our assigned domain, that of finance - mergers and acquisitions in particular."

An important component of PANGLOSS is the PANGLOSS Ontology, a large conceptual network which supports the semantic processing of the other PANGLOSS modules. When it is complete, this network will contain 100,000 nodes representing commonly encountered objects, entities, qualities, and relations. It is being built partly by merging WordNet, the semantic word database based on psycholinguistic principles by George Miller at Princeton, and Longman's LDOCE dictionary... ISI is developing a suite of algorithms which match LDOCE definitions and WordNet definitions in order to flesh out this network. Because they are aiming at broader coverage than has previously been possible in MT systems, part of the group's strategy is to develop automatic and semi-automatic methods of knowledge acquisition for the system. Since dictionaries and corpora are not always perfect sources of knowledge, initially they still check the results.

The PANGLOSS project is one of three MT projects currently being funded by ARPA. Within a three-year programme, it is supporting the statistics-based prototype of Peter Brown's group at IBM, a hybrid approach by Dragon Systems, which has no experience with translation but has booked impressive success in the speech arena, and the CMU/NMSU/ISI triad with its knowledge-based orientation. While Dragon is starting with a clean slate, the IBM and PANGLOSS teams are building on research which has been around for awhile. ARPA would clearly like to see some of this well-cloistered technology get out into the world. ARPA's carrot is its substantial funding of concrete, short-term goals; its stick is its annual evaluations.

.... The next ARPA evaluation commences on May 15. The three MT systems will be tested on a collection of twenty-two newspaper articles, mostly in the domain of mergers and acquisitions. A team of evaluators will give marks for adequacy, style, and comprehension, ranking MT output against machine-aided and fully manual translations not identified as such. When the program is completed after three years... what will happen next? "Hopefully, there'll be a follow up, PANGLOSS II," replies Hovy. Thereafter, parts of PANGLOSS might then be ripe for commercial exploitation in collaboration with an industrial partner - a good five years down the line. Hovy would like to see something of practical import eventually result from the beehive of activity surrounding language processing, but he warns us that we should not expect radical breakthroughs.

REVIEWS and REPORTS

JEIDA publishes MT survey results

John Hutchins

Appearing some months after its report on evaluation methodology [see MTNI#4: 19-20, and item below], the Japan Electronic Industry Development Association has published the English version of the results of a questionnaire sent out in November 1991. The report is entitled *The survey of the current status of research and future trends in machine translation and natural language processing* [for details see 'Publications Received'.]

The questionnaire on MT and NLP was large and detailed – deliberately so. JEIDA did want to produce a superficial survey giving only simplistic answers to difficult issues. In all 2210 were distributed to researchers, manufacturers and experts in 49 countries. JEIDA must surely have been disappointed at the poor response rate; perhaps its ambitions had been too high, and the detail and depth had deterred rather than encouraged respondents. There were in total 173 returns (8.1% of those sent out), with the best response from Japan (16%, or 96 out of 610) - outside Japan, the highest numbers of returns were from Germany (21) and the United States (10). Respondents were predominantly researchers (66%) and/or working in companies (60%). The systems involved were almost equally divided between MT (44%) and NLP (33%).

The questionnaire had two parts: one concerned with individual respondents, the other concerned with features of systems. Given the unrepresentativeness of the systems covered, the second part has unfortunately doubtful value as a picture of the current situation - e.g. translation methods were almost exclusively the 'transfer' type (25); the 'interlingua' design was represented by only 9 responses, and knowledge-based, example-based, corpus-based and stochastic-based translation methods were not represented at all. It is a pity then not to have reliable answers to such questions as: purpose of system (practical/commercial? experimental?), languages involved, translation direction (one-way, two-way, multi-target), language-dependence, domain-dependence, quality of translation aimed for, hardware configuration (workstation, personal computer, mainframe), dictionary sizes (in these responses, over half had dictionaries with more than 50,000 entries), methods for dealing with homonymity, bilingual lexical differences, compounds, use of semantic features, methods of syntactic analysis, size of corpus, and much more. The report reproduces respondents' comments to many questions, and these might prove of value. But probably most useful will be the schematic presentations of the systems for which there were returns, even if coverage is unrepresentative.

There is more, perhaps, to be learnt from the first part of the questionnaire concerned with the opinions of researchers. In background, researchers were – not surprisingly – computer scientists (45%), linguists (29%) or computational linguists (12%), with the rest mainly from other natural sciences (24%), and only 11% with a background in translation. Where did they see the main problems for current MT? For 32% these were 'elementary techniques' (i.e. the basis processes of parsing, disambiguation, semantics), for 19% they were 'systems' (i.e. speed, robustness, environment, human interaction) - although the Japanese respondents rated 'system' problems higher than 'techniques' (28% compared with 24%). Other answers were: 'dictionaries and grammars' (11%), 'linguistic theories' (8%), and 'translation quality' (7%). What then were the areas of study of greatest importance for further development of MT technology? Highest on the list came the dictionary (63% of all respondents), contextual analysis (61%), learning mechanisms (58%), and semantic analysis (56%); then came discourse analysis (49%), knowledge acquisition (45%) and large-scale knowledge bases (45%). Further down came semantic representation (31%), non-grammatical phenomena (29%), syntactic analysis (21%); and translation theory just 16%. Asked to predict the future, respondents thought that MT would be in common use by researchers within 5 years (39%) or between 6-10 years (31%), by translators within 5 years (44%) or between 6-10 years (32%), and by business people within 5 years (35%) or between 6-10 years (35%). But use by the general public was pushed further into the future: within 5 years (12%), between 6-10 years (26%), between 11-20 years (36%). Such optimism may well be typical of the MT community, however unjustified it might appear from the previous history of MT developments. It would be interesting to discover whether this JEIDA survey is a true reflection of what MT researchers as a whole believe.

JEIDA Machine Translation Evaluation Methodology

Hirosato Nomura and John Hutchins

The Japan Electronic Industry Development Association (JEIDA) has been studying the methodology of machine translation evaluation in its Machine Translation Market and Technology Study Committee (Chairman: Hirosato Nomura) for several years. The first version of its evaluation criteria were published in March 1992 in a report written in Japanese entitled "Study on Machine Translation Development". The report consisted of two parts: Study of Machine Translation Evaluation (pp. 17-301), and Survey of Natural Language Processing Technology (pp. 303-660). A short version of the first part was translated into English (127 pages) and was presented at the MT Evaluation Workshop held at San Diego in November 1992.

The English report was the reviewed in *MTNI#4 (p.19-20)*. In this report (*JEIDA Methodology and Criteria on Machine Translation Evaluation*) the committee proposed two sets of criteria: one for users and the other for developers. The criteria set for users is further separated into two parts; one for economical evaluation and the other for technological evaluation. Application of the economical evaluation criteria should enable users to estimate the cost savings from the use of MT in their situation. Application of the technological evaluation criteria should enable users to estimate the technical capabilities of systems. The criteria provide both a check list for evaluation an algorithm for the numerical rating of individual factors, and a visual framework for representing the evaluation results (a radar chart). The committee tested the criteria in simulation studies and presented results in the report. The report included also proposals for evaluation criteria for developers covering all the technical aspects of machine translation: problems of computational linguistic theory, text corpus, lexicon, grammar, parsing, transfer, generation, user interface, system implementation, environment integration, operation, editing, network transmission, etc.

The committee has continued to elaborate the criteria and has published a second report (300 pages) this March. This report will not be translated into English for financial reasons. It presents a field test of the first version of the evaluation criteria and a discussion of the experiments. It also discusses revisions of the criteria and then proposes version 1.2 of the criteria (it is planned to publish the full set of the second version in March 1994.) The 1.2 version consists of three sets of criteria: the first is a set of evaluation criteria for users (for both economical and technological evaluation), the second is a set of evaluation criteria for developers (technical features of computational linguistics, natural language processing, human interface, system integration, etc.), and the third is a set of evaluation criteria for the linguistic quality of translations (distinct from the criteria for developers in the first version.)

Research on Human Language Technology
Joint NSF/DARPA Initiative Announcement

[Edited extract from *Linguist* discussion list]

The US National Science Foundation (NSF) and Defence Advanced Research Projects Agency (DARPA) has issued the following Announcement.

Introduction. Beginning in 1993, the Information, Robotics and Intelligent Systems Division of the Computer, Information Science and Engineering Directorate of NSF and the Software and Intelligent Systems Technology Office of DARPA will support jointly innovative, multi-disciplinary, research projects in the general area of human language technology. The motivation for this initiative is the favourable research environment provided by continuing advances in computer technology. This was recognized and encouraged in the 1992 report of the NSF-sponsored Workshop on Spoken Language Understanding. Computing systems now available and affordable for research are proving adequate to support major advances in natural language understanding, speech recognition, machine translation, and other human language technologies. It is now becoming possible to create realistic computer models of human language mechanisms. We also take account of the synergistic advantage of the combined common research interests of NSF and DARPA in artificial intelligence and human language technology. Therefore, the time is ripe for accelerating efforts in these areas of artificial intelligence.

This NSF/DARPA joint research initiative has the following objectives:

1) To support the long-term goal of achieving effective, general, human-computer communication through the medium of human language.

2) To accelerate progress in the development of the scientific and technical foundations of automatic human language processing by computer.

3) To broaden the scope of research on human language technology by including novel ideas and approaches beyond those now being pursued in ongoing research programs.

4) To facilitate technology transfer by building on NSF's interest in basic science and DARPA's interest in technology and system-level functionality. To this end, industrial/university collaboration is required in the proposed research.

Proposals should be submitted to NSF's Interactive Systems Program. The selection of projects for funding will be made through the normal NSF merit review process with DARPA's participation. Successful proposals will receive support for a three-year period.

Areas of Interest. This initiative is dedicated to the general area of human language technology and, in particular, to aspects of human language understanding. There is special interest on fundamental issues common to different languages and to different communication modalities, and on both language production and language recognition/comprehension. Projects with general technical applicability across various languages and modalities are encouraged.

Human language is an area of empirical study, and carefully designed corpora for research play a key role in the success of a project. Since the creation of such corpora is a costly endeavour we anticipate that prospective investigators needing such data will make full use of existing corpora. The Linguistic Data Consortium (LDC, 441 Williams Hall, U. of Pennsylvania, Philadelphia, PA 215/898-0464, e-mail ehodas@unagi.cis.upenn.edu) is a good source of available corpora supporting research on human language.

Scope of support. This is a one-time solicitation that extends over a three year effort and is expected to provide funding up to $2 million per year. The number and size of awards is contingent on the quality of proposals and the availability of funds. An upper limit of $300,000 per year for three years for research teams of 2-3 researchers is suggested. Awards under this initiative may provide support for principal investigators, graduate students, postdoctoral research associates, specialized equipment and software and databases necessary for the research proposed. Industrial participation is required and collaborative cost-sharing is encouraged. Cost-sharing arrangements must be clearly described in the proposal.

Inquiries. Proposals should be submitted to NSF following the guidelines of the publication NSF 92-89, *Grants for Research and Education in Science and Engineering: An Application Guide*. For technical information, prospective applicants may contact either NSF or DARPA program office: NSF: Dr. Oscar N. Garcia: (202)-357-9554; ogarcia@nsf.gov; Fax: (202)-357-0320. DARPA: Dr. George R. Doddington: (703)-696-2259; doddington@darpa.mil; Fax: (703)-696-2202.

Applications. Fifteen (15) copies of the proposal must be addressed to: Announcement No. NSF-93-19, National Science Foundation - PPU 1800 G Street NW Room 233, Washington DC 20550-0002 and must be received following the guidelines of the publication NSF 92-89 mentioned above by the deadline of May 17, 1993.

EAGLES take off in Europe

[Edited extracts from *EAGLES Information sheet 1*: April 1993]

The integration of natural language and speech processing into complex Information Technology applications has been hampered by the lack of generic technologies and of large-scale language resources. In Europe, there is particular concern as the Language Industries are mainly driven by small and medium-sized companies providing highly customised applications. These are slow to develop and modify, mainly because of the high costs of building the natural language or speech resources required for such applications. An associated problem is the diversity of formats and variable linguistic specificity of existing resources which hinder their re-use and indeed engender duplication of effort. Several European projects and interest groups have been addressing these problems and it has become clear that harmonisation of both linguistic information and its representation is required.

In order to further consolidate consensus-building amongst language researchers, EAGLES -- the Expert Advisory Group on Language Engineering Standards -- was recently launched within the framework of the CEC's DGXIII Linguistic Research and Engineering (LRE) Programme. Started in February 1993, the EAGLES initiative will complement and expand these efforts towards uniform data representation and will strive to improve harmonisation of linguistic information. EAGLES is intended to accelerate the provision of standards for the development, exploitation and evaluation of large-scale language resources, and is expected to produce an initial set of wide-ranging proposals for European *de facto* standards by mid-1995.

The aims of EAGLES are:

‹ to produce public, commonly agreed specifications and guidelines for specific areas of language engineering, based on pooling results from current European efforts and exploiting networks of expertise;

‹ to complement European R&D projects;

‹ to promote the adoption of EAGLES results in future R&D ventures; and

‹ to feed results to national and international standardisation initiatives.

Given the scope of the LRE programme, EAGLES will select only certain key issues of relevance to the standardisation process, and will initially be able to involve only a selection of relevant European and national research and development activities.

The structure of EAGLES has resulted from recommendations made by leading industrial and academic centres, and by the CEC's Language Engineering strategy committees. More than 30 research centres, industrial organisations, professional associations and networks across the EC are donating labour towards the effort. With financial support from the CEC's LRE Programme, coordination is being carried out by the Consorzio Pisa Ricerche (Italy).

EAGLES consists of five Working Groups, hosted by designated R&D centres; a Management Board; and a central support team.

The definition of specifications and guidelines is carried out in the Working Groups. These are concerned with common methodologies for the following five areas:

Text corpora - host: Instit.Cervantes, Madrid, Spain - chair: A.Zampolli (I)

Computational lexicons - host: GSLI-ERLI, Paris, France - chair: M.Nossin (F)

Grammar formalisms - host: DFKI, Saarbrücken, Germany - chair: H.Uszkoreit (G)

Evaluation and assessment - host: Centre for Sprogteknologi, Copenhagen, Denmark - chair: M.King (CH)

Spoken language - host: Logica Cambridge Ltd., Cambridge, UK -chair: R.Moore (UK)

Each Working Group is made up of experts from European industry and academia and is hosted by a well-known private or publicly sponsored research laboratory which offers basic logistics and scientific-technical support.

The Management Board is charged with overall coordination and administration; it provides a forum for scientific exchange among EAGLES organisations; it supervises the Working Groups; and endorses and promotes their results.

The Board (Chair: Prof Rohrer, Stuttgart University; Vice-Chair: Dr Peckham, Logica Cambridge Ltd) is at present composed of 13 organisations representing European projects in Natural Language and Speech (ACQUILEX, DELIS, EUROLANG, GENELEX, GRAAL, MULTILEX, NERC, ONOMASTICA, PLUS, RGR, SAM-A, SUNDIAL, and TWB) and several European associations and co-ordinating bodies (EACL, ELSNET, ESCA and FOLLI). Private research laboratories and Information Technology providers specialised in language engineering account for almost 50% of the Board's membership. Other groups may apply for membership if they represent multinational, multilingual European R&D language engineering efforts and are able to commit free, competent labour.

The central support team (Director: A.Zampolli, CPR, Pisa) is responsible for ensuring cohesion of the distributed activities, for mediating between the participating bodies, and for supplying administrative/technical support.

Each Working Group has identified a number of Subgroups to address specific topics. The members of these are drawn from relevant R&D projects initiated under the main European programmes or from organisations with relevant experience.

User groups including national and EC research teams will be consulted both by Working Groups and by a system of 'Affiliated Projects' in an attempt to reconcile pre-normative research issues with user requirements. EAGLES will collaborate with international groups involved in standardisation, notably the TEI, and will explore collaborative arrangements with data collection efforts in the USA such as the LDC. It will also interact with LRE-sponsored resource management and dissemination networks such as NERC and the planned RELATOR project.

Under the editorship of N.Calzolari (CPR, Pisa) and J.McNaught (UMIST), the results of the EAGLES initiative will be published and widely disseminated as a set of 'Guidelines and Common Specifications' by the second quarter of 1995. It is hoped that work will continue in future Community R&D actions in the language technology field.

For further information contact: Ms Tarina Ayazi, EAGLES Secretariat. Fax: +39-50-589055. Email: eagles@icnucevm.cnuce.cnr.it

Bibliographical Database On Computers, Linguistics and Communications

[Extract from the *Langage Naturel* electronic discussion group]

For the last 15 years, the University of Montreal has been compiling a bibliographical database on all aspects of computer processing of natural language communications. The bibliography, which now holds more than 67,000 references, is indexed with a thesaurus of over 3,400 keywords. More than 13,000 titles are related to artificial intelligence.

The references cover the period beginning with the inception of the computer to the present and include theses, research reports, books, articles from specialized periodicals, papers in conference proceedings, etc. The entries were obtained mostly by systematically scanning more than 400 periodicals and 800 conference proceedings.

Some of the thematic sections of the database are near completion and will be published in print in the coming months. Each thematic volume will have a two-level analytical index.

Many researchers collaborated by sending us their lists of publications. All others who are interested are invited to do so.

In the list that follows, the numbers refer to the approximate number of entries of some of the subsections of the database.

Natural language interfaces (3000), Text understanding (3800), Parsing (7000), Computational morphology (2000), Text generation (2000), Speech analysis, coding and synthesis (2800), Speech recognition and understanding (3000), Text information extraction (2000), Information retrieval (3000), Computer translation (7000), Mathematical and formal linguistics (3000), Cognitive linguistics and psycholinguistics (1600), Literary computing (3000), Quantitative and statistical linguistics (2400), Computer assisted language teaching (5500), Electronic document processing (2300), Computational lexicography (3000), Optical character recognition (2900), Character processing (2200), Communicating through computers (2100), Corpus linguistics and dialect study (1000).

For more information about this database, contact Conrad F. Sabourin, P.O. Box 187, Snowdon, Montreal, Qc H3X 3T4, Canada; email: sabourco@ERE.UMontreal.CA

FROM THE ARCHIVES...

The first MT patents

John Hutchins

Today machine translation means using a computer to translate natural languages. But it was not always so. The first suggestions that languages could be translated mechanically were made before electronic digital computers were even dreamt of. While the real history of MT began in March 1947 with correspondence between Warren Weaver and Norbert Wiener and with tentative discussions between Weaver and Andrew D. Booth (not in 1946 as Booth himself sometimes asserted later), it is nevertheless legitimate to recognise precursors in two patents submitted sixty years ago in 1933. One patent was issued in Paris on 22 July 1933 to Georges Artsrouni; the other was issued in Moscow on 5 September 1933 to Petr Petrovich Troyanskii. Both patents referred essentially to the construction of mechanical multilingual dictionaries.

Of course, the idea of translating 'mechanically' via dictionaries was itself not new – the essential ideas can be traced back to Descartes and Leibniz, who both proposed numerical codes to mediate between languages, and specific proposals for mechanical dictionaries were published from the middle of the 17th century onwards (e.g. Athanasius Kirchner 1663 and Johann Joachim Becher 1661) together with various proposals for universal languages (e.g. the famous Real Character of Bishop John Wilkins in 1668). A little later, in 1726, the idea of mechanizing language was widely popularized by Jonathan Swift in his Gulliver's Travels, where in the Academy of Lagado, Gulliver encountered scientists working on a machine for text generation, consisting of "bits of wood covered on every square with paper pasted on them, and on these papers were written all the words of their language, in their several moods, tenses, and declinations,..." It was not, however, until this century that specific devices were actually constructed for automating translation.

An account of Artsrouni's patent has been given by Michael Corbé ('La machine à traduire française aura bientôt trente ans' *Automatisme* 5 (3), 1960, 87-91). Artsrouni was a French engineer of Armenian origins and a former student of one of the major schools of St Petersburg. Apparently under development since 1929 the device, which Artsrouni called a *cerveau mécanique* (anticipating the popular name for the first electronic computers), was intended for a wide variety of tasks: the production of timetables and telephone books, for accounting, and for deciphering and encrypting messages. In 1937 the device was exhibited with great interest at the World Exhibition in Paris. A number of state organisations signed contracts with the inventor; the postal service, for example, ordered one to deal with money orders, and the railway service intended to use it for printing tickets. Only the war and the occupation of France prevented these plans coming to fruition.

The idea of using his invention for translation was prominent from the outset. In the 1933 description, Artsrouni explained that his 'brain' could "translate from one foreign language into any one of three other languages registered", and indeed was not limited to four languages or to a restricted number of words. In essence, the device consisted of four components: a memory (called *bande des réponses*), a keyboard for entering words (*tête de lecture*), a search mechanism (*sélecteur*), and an output mechanism (*sortie*). The core component was the 'memory', a paper tape 40 cm wide and up to 40 meters in length stored on two rollers and moved by cogs on lateral perforations (like a photographic film roll). Dictionary entries were recorded in lines of four columns, one column for each language; up to 40,000 lines were possible, i.e. 40,000 lexical items for four languages. In fact, Artsrouni suggested that the number could be doubled by having entries in two colours (red and blue) on each line. The user was able to modify and add entries as necessary, since the order of items was quite free. The memory was searched by entering sought words at the keyboard which was linked to the 'selector' - another tape (paper or metal) on two rollers -containing codes for all entries in the memory. Output was displayed in a series of windows on the keyboard. Operation of the mechanical dictionary consisted, therefore, of moving the selector tape to match the input word and simultaneously moving the memory tape so that the translation could be read at the windows on the keyboard. In the earliest model it was claimed that the selector and memory tapes could move through the 40,000 lines in 60 seconds; in later models the speed was reduced to an average of 3 seconds.

Corbé compared Artsrouni's device to the machine constructed by Gilbert King for IBM, based on a device called the 'photoscopic store'. This too was essentially little more than a mechanical dictionary, although promoted as an MT system (and in fact used by the US Air Force from 1959 to 1964). In as much as his device was essentially the same idea, Corbé felt justified in claiming Artsrouni as a precursor of MT. Today we are rather less inclined to refer to automated dictionaries as translation systems.

In the case of Troyanskii, however, we do have a genuine precursor of machine translation. In the preface to a collection of his papers published in 1959 (*Perevodnaya mashina P.P. Troyanskogo*. Moskva: Akademiya Nauk SSSR, 1959) we read the following biographical details:

"Petr Petrovich Troyanskii was born in January 1894 in the family of a railway repair-shop worker in Orenburg. The family had 14 children and the living was hard. P.Troyanskii finished a parish school in Orenburg and passed gymnasia examinations without attending classes, after which he entered the University of St. Petersburg. He made his living by giving lessons. World War I prevented P.Troyanskii from finishing the university. After the Great October Revolution he entered the Institute of Red Professorate. Afterwards he taught social sciences and the history of science and technology at higher educational establishments. He also participated in compiling the Technical Encyclopedia and the Great Soviet Encyclopedia. In those years he devoted more and more time to putting into practice his idea of a translating machine. A serious illness - stenocardia -prevented P.Troyanskii from completing work on mechanizing translation which he considered the cause of his whole life. Petr Petrovich Troyanskii died on the May 24, 1950." [Translation by Evgenii Lovtskii]

Troyanskii's patent "for the selection and typing of words while translating from one language into another" consisted of a sloping table on which could be moved freely in all directions a broad band comprising a multilingual dictionary of entries arranged, like Artsrouni's, in columns. In Troyanskii's case the entries were not full word forms but stems (e.g. infinitives of verbs). Troyanskii envisaged three stages. In a 'pre-editing' stage a user knowing only the source language identified stems and endings, and replaced the latter by pre-defined 'logical forms'. In the second, purely mechanical, stage the entries for source word-stems were located, the corresponding target words were photographed onto a tape and, at the same time, the 'logical forms' were typed out [see the attached translation of the patent for details]. In a 'post-editing' stage a user knowing only the target language provided the morphologically correct target forms.

For his logical forms, Troyanskii borrowed from Esperanto: nouns in the nominative were given endings in –o, plural forms in –j and oblique cases were indicated by –n; adjectives have the ending –a, verbs in the present tense end in –as and infinitives in –i. The following extract from the 1959 collection illustrates translation from the French sentence *Le parti périt, s'il commence à cacher ses erreurs* into Russian:

Le parti–o	partiya–o	Partiya
périr–as	pogibat'–as	pogibaet
si	esli	esli
il	on	ona
commencer–as	nachinat'–as	nachinaet
cacher–i	skryvat'–i	skryvat'
son–ajn	svoi–ajn	svoi
l'erreur–ojn	oshibka–ojn	oshibki

Various improvements were made to the device in later years, and by 1941 an experimental machine was operational, similar in a number of respects to the Harvard Mark I machine, developed between 1938 and 1942, which is regarded as the direct forerunner of ENIAC, the first computer. But support in the Soviet Union was not forthcoming, and his proposals remained unknown even in his own country until MT research had been underway for more than a decade.

Although his patent described only the operations of the mechanical dictionary, Troyanskii stressed in his writings the belief that all stages of translation could be automated. In this respect, he anticipated some central MT concepts, proposing the now familiar three-stage model of analysis, transfer and synthesis, and advocating the use of 'quasi-logical' interlingual elements. The linguistic details were not worked out, however; there was no discussion (or perhaps even awareness) of the problems of treating idiomatic expressions, homonyms or differences of word order. Nevertheless, there is little doubt that if the electronic computer had been available for the realization of his ideas Troyanskii would today be widely regarded as the true 'father' of machine translation.

Troyanskii's patent

[Translated by *Evgenii Lovtskii*]

INVENTION AUTHORSHIP CERTIFICATE

Description of a machine for selecting and typing words when translating from one language into another or several others simultaneously.

To P.P.Troyanskii's claim of invention and certificate
issued on September 5, 1933 (priority no. 134430)
The fact of granting an authorship certificate to P.P.Troyanskii
was made public on January 31, 1935.

The machine in question is designed for selecting and typing words when translating from one language to another or several others simultaneously and essentially consists of a special perforated belt to which are affixed words in different languages. The belt can move over a desk; the perforations are used for positioning it in front of a photographic camera, adjacent to which is located a typewriter with additional keys for typing conventional signs alongside the photographed word. A view of the machine is given in the drawing.

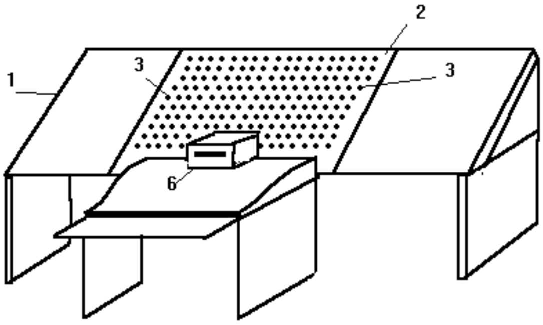
As is shown in the drawing, the machine consists of a smooth sloping desk (1), over which a belt (2) can move easily and freely in different directions. The belt has perforations (3) for pins which position it against aperture (6). A six-language (or any other number of languages) parallel dictionary is arranged alphabetically on the belt's surface in such a way that words beginning with more frequent letters (e.g. K, M, P, etc.) are located closer to the middle.

The machine is operated in the following way. According to the word being translated, the belt is moved to a position where the word to be translated or a row of words in different languages find themselves against aperture (6), then the belt is stopped and the word is photographed on a light-sensitive film. At the same time conventional signs for logical parsing are typed on a paper tape, then the film and the paper tape are moved one line forward and the belt (2) is re-positioned for processing the next word or a row of words when translating into several languages, and so on.

From the translation produced in this manner and recorded on the film and the tape, fixed or glued together, with columns of photographed words and typed signs of logical parsing, a typist types a coherent text which goes to a special reviser who imparts to the words appropriate forms according to the conventional signs of logical parsing. After re-typing the text goes to a literary editor for final editing.

INVENTION SUBJECT MATTER

A machine for selecting and typing words when translating from one language into another or several others simultaneously, characterized by the use of a circular belt (2), with words in different languages pasted to it and with perforations (3) for positioning the appropriate word or words against an aperture in the desk above which a photographic camera is arranged for recording selected words on a light-sensitive film and, nearby, a typewriter with additional keys for typing conventional signs on a paper tape alongside the photographed word.



Expert and editor A.G.Bremzen

PUBLICATIONS RECEIVED

Journals

Language Industry Monitor, *no.13 (Jan-Feb 1993).* p.1-4: Making MT work [on Lexi-tech (Canada) using Logos to translate shipbuilding documentation] - p.12: A report on MT from the ATA [by Christine Miller]; *no.14 (Mar-Apr 1993).* p.10-11: Pangloss: interlingua vivat? [see item in this issue.]

Language International, *vol.5 no.1 (February 1993).* p.29-31: Translating in America, MT evaluation and new frontiers (Bob Clark) [report of AMTA San Diego meeting, November 1992]
vol.5 no.2 (April 1993). p.5-6: Technology and language [report of CEC symposium 'Technology and Language, Europe 2000', London 15 January]. - p.8-9: Systran, the Telinfo contribution. - p.14: Translating and the Computer [report of 14th conference, November 1992]

Machine Translation, *vol.7, no.1-2, 1992*: Special issue on text generation.
Contents: p.1-4: Guest editor's note (Richard Kittredge). - p.5-40: Towards meaning-based machine translation: using abstractions from text generation for preserving meaning (John A.Bateman). -p.41-60: Utterance generation using conceptual representation (Eric Bilange, Jean-Marie Lancel, Miyo Otani). - p.61-98: Interactions between linguistic constraints: procedural vs. declarative approaches (Martin Emele, Ulrich Heid, Stefan Momma, Rémi Zajac). - p.99-124: Text planning with opportunistic control (Sergei Nirenburg). - p.125-134: Book review: C.Paris et al eds. Natural language generation in artificial intelligence and computational linguistics (William J.Black)

Meta: Journal des traducteurs, *vol.37, no.4, décembre 1992.* Numéro spécial sous la direction de Monique C.Cormier et Dominique Estival. Etudes et recherches en traductique, Studies and researches in machine translation. Presses de l'Université de Montréal. (ISBN 2-7606-2439-0)
Contents: p.583-594: User driven development: METAL as an integrated multilingual system (Thomas Schneider). - p.595-609: Télécommunications et micro-informatique, les alliés du traducteur d'aujourd'hui: Systran s'adapte... (Claude Bureau). -p.610-623: La traduction automatique au service de l'utilisateur monolingue (Laurence Jacqmin) [on Babel-R, Brussels]. - p.624-634: TRADEX, un système de traduction del télex (Jean-Marc Aumaitre, Laurence Horel, Jean-Marie Lancel) [at Cap Gemini Innovation, Paris]. - p.635-646: Problèmes de traduction automatique dans le sous-langage des bulletins d'avalanches (Pierre Bouillon, Katharina Boesefeldt) [at ISSCO, Geneva]. -p.647-656: La génération de textes multilingues par un utilisateur monolingue (Harold Somers, Danny Jones) [at UMIST, Manchester]. - p.657-680: Unification and machine translation (Louisa Sadler, Doug Arnold). - p.681-692: Simplifying the complexity of machine translation (Randall Sharp, Oliver Streiter) [on CAT2 at Saarbrücken]. - p.693-708: ELU, un environnement d'expérimentation pour la TA (Dominique Estival) [at ISSCO, Geneva]. - p.709-720: Tools for machine-aided translation: the CMU TWS (Sergei Nirenburg). - p.721-738: La bi-textualité: vers une nouvelle génération d'aides à la traduction et la terminologie (Pierre Isabelle) [at CCRIT]. - p.738-760: La prétraduction automatique, outil de productivité et d'évolution professionnelle (Claude Bédard). - p.761-769: La technologie langagière au Secrétariat d'État du Canada: une réalité quotidienne (Klaire Tremblay). - p.770- 790: Aide au transfert lexical dans une perspective de TAO: expérimentation sur une lexique non terminologique (Ariette Attali et al.) [at CELTA, Nancy]. - p.791-801: Is translation symmetric? (Louis Des Tombe). - p.802-816: Machine translation research in Czechoslovakia (Jan Hajič, Eva Hajičova, Alexandr Rosen). - p.817-827: L'évaluation des systèmes de traduction automatique dans le cadre d'un service de traduction (Margaret King). - p.828-846: La traduction automatique: l'ordinateur au service des traducteurs (Brigitte Roudaud) [on B'VITAL, SITE and Eurolang]

Books

The survey of the current status of research and future trends in machine translation and natural language processing. Tokyo (Japan): Japan Electronic Industry Development Association, December 1992. 212+38 pp. [See report in this issue.]

Building lexicons for machine translation. Papers from the 1993 Spring Symposium (Technical report SS-93-02). Menlo Park, Cal.: AAAI Press (American Association for Artificial Intelligence), 1993. ISBN: 0-92980-39-3. [Available from: AAAI Press, 445 Burgess Drive, Menlo Park, California 94025.]
Contents: p.1-9: Three-level knowledge representation of predicate argument mapping for multilingual lexicons (Chinatsu Aone & Doug McKee). - p.10-11: Large lexical European projects and the multilingual aspect (Nicoletta Calzolari & Antonio Zampolli). p.12-21: Multilingual lexical representation (Ann Copestake & Antonio Sanfilippo). - p.22-23: Tracking verbs across languages (Beth Levin). - p.24-33: How to overcome translation mismatches - an inference driven mapping between meaning representations (Bianka Buschbeck-Wolf). - p.34-42: Lexical issues in dealing with semantic mismatches and divergences (James Barnett & Elaine Rich). - p.43-53: Constraints on the space of MT divergences (Bonnie Dorr & Clare Voss). - p.54: Grammatical semantics and multilinguality: what stands behind the lexicon? (John A. Bateman). - p.55-56: Using on-line thesaurus in machine-aided translation systems (Sylvie Regnier, Frederique Segond & Shirley Thomas). - p.57: The status of semantics in Multilex (Nicole Modiano). - p.58: The MT lexicon and the translation of compounds (Paul Bennett, Marta Carulla & Kerry Maxwell). - p.59-62: A probabilistic approach to Japanese lexical analysis (Virginia Teller). - p.63-68: The use of machine-readable dictionaries in the Pangloss project (Stephen Helmreich, Louise Guthrie & Yorick Wilks). - p.69-80: Collocations, dictionaries, and MT (Dirk Heylen, Kerry Maxwell & Susan Armstrong-Warwick). - p.81: Towards a theory of polysemy (D.Alan Cruse). - p.82-85: Translation by confusion (Hinrich Schuetze). - p.86-92: Combining logic-based and corpus-based methods for resolving translation mismatches (Megumi Kameyama, Stanley Peters & Hinrich Schuetze). - p.93-97: Automatic dictionaries in the ETAP-3 system (D.Ju. Apresjan, et al.). - p.98: Get it where you can: acquiring and maintaining bilingual lexicons for machine translation (Mary S. Neff et al.). - p.99: Inflectional morphology needs to be authenticated by hand (Robert L.Mercer). - p.100: Using machine readable dictionaries for the creation of lexicons (David Farwell, Lousia Guthrie & Yorick Wilks). - p.101: Merging LDOCE and WORDNET (Kevin Knight). - p.102-113: A producer-consumer schema for machine translation within the PROLEXICA project (Patrick Saint-Dizier). - p.114-121: The semantic and stylistic differentiation of synonyms and near-synonyms (Chrysanne DiMarco, Graeme Hirst & Manfred Stede). - p.122-131: Principles and idiosyncrasies in MT lexicons (Lori Levin & Sergei Nirenburg). - p.132-135: What kind of information is necessary for NLP and MT? (Makoto Nagao).

Sixth Conference of the European Chapter of the Association for Computational Linguistics. Proceedings of the conference, 21-23 April 1993, OTS - Research Institute for Language and Speech, Utrecht University, Utrecht, The Netherlands. ACL: 1993.
Contents include: p.12-20: Experiments in reusability of grammatical resources (Doug Arnold et al.). - p.113-119: Automating the acquisition of bilingual terminology (Pim van der Eijk). - p.149-157: Rule-based acquisition and maintenance of lexical and semantic knowledge (Donna M.Gates & Peter Shell). -p.193-202: Restriction and correspondence-based translation (Ronald M.Kaplan & Jürgen Wedekind). - p.449-453: Text alignment in a tool for translating revised documents (Hadar Shemtov). -p.466: Long sentence analysis by domain-specific pattern grammar (Shinichi Doi et al.). - p.468: The Pangloss Mark I MAT system (Robert Frederking et al.). - p.476: ITS-2, an interactive personal translation system (Eric Wehrli & Mira Ramluckun).

First Conference of the Pacific Association for Computational Linguistics. Proceedings of the Conference. 21-24 April 1993, Simon Fraser University, Vancouver, British Columbia, Canada.
[Available from: Paul McFetridge, Department of Linguistics (email: mcfet@cs.sfu.ca), or Fred Popowich, School of Computing Science (email: popowich@cs.sfu.ca), Simon Fraser University, Burnaby, British Columbia, Canada V5A 1S6.]

Contents include: p. 288-296: Translation of metonymy in an interlingual MT System (Takahiro Watao & Stephen Helmreich). -p. 297-303: Choosing the right word: lexical knowledge and context in machine translation (John Phillips). - p. 304-308: Tuning of a machine translation system to wire-service economic news (Teruaki Aizawa, Naoto Katoh & Masoko Kamata) [at NHK]. -p. 304-313: The integration of MT and MAT (Robert Frederking, Dean Grannes, Peter Cousseau & Sergei Nirenburg) [at CMU].

Items for inclusion in the 'Publications Received' section should be sent to the Editor-in-Chief at the address given on the front page. We rely on readers to notify us of their publications.