MACHINE TRANSLATION REVIEW

The Periodical
of the
Natural Language Translation Specialist Group
of the
British Computer Society

Issue No. 1
April 1995

The Machine Translation Review incorporates the Newsletter of the National Language Translation Specialist Group of the British Computer Society and appears twice yearly.

The Review welcomes contributions, articles, book reviews, advertisements, and all items of information relating to the processing and translation of natural language. Contributions and correspondence should be addressed to:

Derek Lewis,
The Editor,
Machine Translation Review,
Department of German,
Queens Building,
University of Exeter,
Exeter,
EX4 4QH
UK

Tel: (01392) 264330
Fax: (01392) 264377
E-Mail: D.R.Lewis@exeter.ac.uk

The Machine Translation Review is published by the Natural Language Translation Specialist Group of the British Computer Society.

ISSN 1358-8346

Contents

Editorial

Welcome to the first issue of the Machine Translation Review. The Review incorporates the Newsletter of the Natural Language Specialist Group of the British Computer Society and aims to continue and develop the Newsletter's well established tradition of informing its members about its activities and about developments in machine translation and related areas. The new title, the improved format, and the twice yearly appearance of the Review are intended to encourage a wide range of contributions. Topics covered in the present issue include the Groups's MNLP Project, a summary of the Cranfield Conference, a description of an MT system under development, and the role of MT in a commercial environment; there are also book reviews and an introduction to linguistic resources on the Internet. Articles (either of an academic or non-academic nature), reviews, advertisements, items of information, letters, and suggestions are warmly welcomed. If you develop, teach, or in any way use software that touches on the concerns of multilingual text processing, please consider contributing to the Review. Arrangements are also in hand to publish the Review electronically alongside the printed version.

Derek Lewis

Group News and Information

Letter from the Chairman

There was some relief when we found that the MT Conference in November last year was going to be well attended and then pleasure when in the event it was such a success. This was thanks to all the hard work of the organising committee, particularly those at Cranfield led by Douglas Clarke. We must also thank the speakers for all their work in preparing and presenting such a varied and interesting set of papers.

The Group's programme of talks in 1994 unfortunately suffered but we hope we can start making up for that soon.

Last year we were delighted to welcome a number of additions to the Committee: Derek Lewis from Exeter University as Newsletter Editor, with the assistance of Catharine Scott from the University of North London, and Tania Reynolds; Ian Thomas as Treasurer; and Roger Harris.

However, we were sorry to lose Frank Knowles, David Robinson and Agnes Kukulska-Hulme, but we thank them for their contribution and their interest in the work of the Group.

I hope you like our new look publication, *Machine Translation Review*, which Derek Lewis has master-minded. We hope you will now feel more inclined to submit articles and reports for publication. We very much want to increase the quantity, quality and range of its contents for your information.

Most of what I said last year about supporting the Group still applies. If you are interested in MT and/or MNLP please try to become more involved with the work of the Group and learn more by taking part. We could still use a Membership Secretary, a Universities and Projects Monitor, and a Librarian to keep in touch with the BCS Library and Publishers in order to report relevant new books.

If you would like any further information about these jobs, such as the numbers involved or how much time they might take, please let me know. If you have any other ideas about how you would like to contribute to our knowledge and your own at the same time, please let me know.

David Wigg

The Committee

The telephone numbers and E-Mail addresses of the Officers of the Group are as follows:

David Wigg (Chair):	01732 455446 (Home) 0171 815 7472 (Work) wiggjd@vax.sbu.ac.uk

Monique L'Huillier (Secretary):	01276 20488 (H) 01784 443243 (W) m.lhuillier@vax.rhbnc.ac.uk
Ian Thomas (Treasurer):	0181 464 3955 (H) 0171 379 9800 (W)
Derek Lewis (Editor):	01404 814186 (H) 01392 264330 (W) 01392 264377 (fax) d.r.lewis@exeter.ac.uk
Tania Reynolds (Assistant Editor):	01444 416012 (H)
Catharine Scott (Assistant Editor):	0181 889 5155 (H) 0171 607 2789 X 4008 (W) c.scott@unl.ac.uk
Ian Kelly (Rapporteur):	01276 857599 (H) idkk@gsi.fr

BCS Library

The following books kindly donated by members have recently been passed to the BCS Library at the IEE, Savoy Place, London, WC2R 0BL, UK, and will be added to the slowly growing number of books on MT. Reference books may be examined in the library. Members of the BCS may borrow non-reference books from this library, either in person or by post. All they have to provide is their membership number. The Library is open Monday to Friday, 9.00 to 5.00 pm. (Tel: 0171 240 1871; Fax: 0171 497 3557).

Sparck Jones, K. and Wilks, Y. (eds) (1985) *Automatic Natural Language Parsing*

Nirenburg, S. (ed.) (1987) *Machine Translation*

Tresman, Ian (1994) *Multilingual PC Directory* (3rd edition)

Books

The following MT or NLP related books have been seen or advertised in the last year. This list does NOT purport to be a list of recommendations. Please let us know of any others you think we should know about, or of any similar books proposed for 1995/96. Reviews of books submitted by publishers are in separate section of this periodical (see page 32).

Atkins, B. T. and Zampolli, A. (eds) (1994) *Computational Approaches to the Lexicon,* OUP: Oxford. Hardback £55.00. ISBN 0-19-823979-3.

Tresman, I. (1994) *The Multilingual PC Directory*, 3rd edition, Knowledge Computing, 9 Ashdown Drive, Boreham Wood, Herts. WD64LZ, UK. £35 + postage. ISBN 1-873091-03-6.

Lipkin, B. S. (1994) *String Processing and Text Manipulation in C,* Prentice Hall. Paperback £27.82. ISBN 0-13-121443-8.

Roach, P. (editor) (1992) *Computing in Linguistics and Phonetics. Introductory Readings*, London: Academic Press. Paperback £14.95. ISBN 0-12-589340-X

Arnold, D., Balkan, L., Lee Humphreys, R., Meijer, S., and Sadler, L. (1994) *Machine Translation. An Introductory Guide*, Oxford: NCC/Blackwell. Hardback £40.00. ISBN 1-85554-246-3. Paperback £18.99. ISBN 1-85554-217-X.

Jensen, K., Heidorn, G. E. and Richardson, S. D. (1993) *Natural Language Processing: the PLNLP Approach*, Boston: Kluwer Academic Press. Hardback £80. ISBN 0-7923-9279-5.

Multilingual Natural Language Processing (MNLP) Project

by

David Wigg

In Newsletter 21 of April 1993 Douglas Clarke suggested that there was a need for some commonly required NLP programs which should be freely available to avoid programmers continually re-inventing the proverbial wheel, and that a sub-group should be set up to see what we could do about it.

Some meetings of people interested in supporting this proposal were held and it was decided that:

* a comprehensive list of MNLP functions should be established first before trying to write NLP applications such as morphological analysis, multilingual word-processing, CALL programs, etc;

* a specification of the required software should be published anonymously by the Group in the so called Public Domain with copyright held by the BCS;

* individuals would then be free to write procedures to meet this specification in a variety of languages, which the Group could also publish.

* these procedures could then be used to write linguistic applications which, being based on virtually universally usable program code and data, could then be used by anyone.

In the last Newsletter (22) I therefore proposed the following project (now expanded in detail) for the Group.

Common Procedures and Common Files

Objectives:

* design general purpose text processing software functions with a system of integrated data/file structures, so that a single all embracing system for text processing may be constructed;

* specify and publish the functions, so that the modules could be written independently in a variety of programming languages by anybody interested in doing so;

* the system should be designed so that program source code could as far as possible be sharable between users of most of the commonly used procedural programming languages, so that users would be able to use their normal language of choice and would not have to learn another computer language when writing linguistic applications;

* a common file structure should also be defined to enable data files to be written to a common system standard, using techniques similar to SGML or TEI coding, so that expensively constructed files would be available to any other user of this software using any computer language on any computer, without the need for adaptation or conversion;

* the Group should publish and make available source code at cost of any versions of the software meeting adequate quality standards, so that the software would be freely available, and the reputation of the BCS and the Group would be maintained;

* the functions and file structures should be easy to use, so that they would be accessible to the widest possible audience and not confined to so-called expert programmers only.

Benefits:

* sharing reusable resources allows individuals to add to the work of others enabling resources to be accumulated instead of dissipated;

* the development of a community of experts with a common 'computer language' and common 'data/file structures';

* the creation of universally sharable data files;

* to enable linguistic specialists to develop computerised linguistic applications more easily.

Unfortunately we have not been able to do much work on this project in the last year, mainly due to the work involved in preparing for the MT Conference in November 1995.

A sub-group was convened last year from about 50 members interested in developing this project, but with such a widespread membership of the Group only a small minority were able to attend meetings in London.

With this in mind I would like to think that any individual, or group, should be able to develop a proposal to meet the objectives outlined above, for the committee to consider.

Sub-group meetings should continue to be held with as many people attending as possible to discuss individual proposals but any reasonable proposal should be supported by the Group by paying for some copying and distribution so that others interested in the project but unable to attend committee meetings can offer constructive criticism or, in the extreme, propose an alternative system.

You probably know that there is a proposal in existence which has already been circulated in an early form and discussed by a small sub-group.

I hope this existing proposal can be further developed but I realise that others may have more time and expertise to produce a better result more quickly, so I suggest that system specifications could be proposed on an individual basis for the Committee to consider. If a developer was a member of the Committee then, of course, he/she would not have a vote involving its acceptance.

It might even be possible for more than one system to be accepted if they were sufficiently different and each had substantial advantages of its own.

Members of the original group of about 50 should have received an up-to-date version of the existing specification recently with a questionnaire. If you have not received these recently but are interested in the project, please let me know so that I can send them to you and put you on the list for the future.

MACHINE TRANSLATION — TEN YEARS ON: CRANFIELD CONFERENCE REPORT 12-14 November 1994

by

Derek Lewis

Introduction

The International Conference, *Machine Translation Ten Years On*, took place at Cranfield University, 12-14 November 1994. The occasion was the tenth anniversary of the previous conference on MT, also held at Cranfield. The 1994 conference was organised by Cranfield University in conjunction with the Natural Language Translation Specialist Group of the British Computer Society. The full conference proceedings are scheduled to appear later in 1995 (for further information, please contact the editor).

Apart from detailed descriptions of prototype systems, the conference provided overviews of general developments in the field of MT. Considerable research is taking place into speech recognition and dialogue systems, and into incorporating features of spoken language and discourse into computer representations of natural language. At the same time, more sophisticated techniques for the statistical analysis of text corpora are emerging that may fundamentally alter the direction of MT research. It is clear that knowledge-based systems representing conceptual information for particular subject domains independently of specific languages are seen as a practical way forward for MT. Another promising direction is the emergence of interactive systems that can be used by non-translators working within a distributed processing environment. Moving away from research and development, the conference afforded practical insights into a number of operational systems. These ranged from large, established systems such as SYSTRAN, to smaller interactive programs for PC. The evaluation and commercial performance of MT systems remains a key issue, alongside the wider question of who actually uses MT. This aspect was addressed by Veronica Lawson, who summarised the results of a recent IAMT survey on the users of MT.

The last ten years and future prospects

In a keynote address, Yorick Wilks gave an update on the state of the art in MT and reviewed the status of the major international projects currently in progress. The 'statistical turn' in MT research and development has highlighted the importance of large-scale resources and reinforced the role of evaluation. International co-operation has grown, which has implications for the role of interlinguas in MT models. This is especially so in Japan and the US, where such models remain popular. There are also implications for evaluation, which can become either a co-operative or a 'hegemonic' exercise. Although co-operation can work successfully, it is proving difficult to share resources effectively.

John Hutchins (University of East Anglia) reviewed research methods and system designs during the decade 1984-1994. By 1984, MT had re-established itself after the negative ALPAC Report of 1966. Apart from SYSTRAN and METEO, which had been operational for some time, the first commercially available systems appeared (ALPS, WEIDNER). Existing research groups (GETA, SUSY, METAL) were supplemented by Japanese teams and notably by the EUROTRA project.

From 1984 to 1989 researchers focused on applying linguistic rules to MT; that is, rules for syntactic analysis and generation, for lexical representation and transfer, and for morphology. Transfer systems predominated (ARIANE, METAL, SUSY, MU, EUROTRA), although interlingual systems appeared in the late 1980s (DLT, ROSETTA). Since 1985 a group at Carnegie Mellon University has been applying a knowledge-based interlingual strategy towards MT which aims to extend the purely linguistic approach into an 'understanding' of a real-word domain (KBMT89, KANT, CATALYST). The knowledge takes the form of a database of semantic networks, from which the system constructs propositions and other textual information (in particular, anaphoric links and topic-comment relationships).

Rule-based research has continued into the 1990s. The EUROLANG team in Grenoble aims to develop a ten language-pair system; the project is based on previous work on ARIANE, on the defunct EUROTRA programme, and on METAL. Other systems include CAT2 (Saarbrücken, based on EUROTRA), PATRANS (Denmark), LMT (IBM), ULTRA (New Mexico), UNITRAN and PANGLOSS (US and New Mexico). At the same time, a new, corpus-based direction has emerged. CANDIDE (IBM) statistically analyses huge volumes of bilingual texts in order to predict translation correspondences: no linguistic rules as such are employed. Similarly, 'example-based' MT relies on memorising pre-translated phrases derived from bilingual text corpora.

During the 1980s rule-based research moved away from mapping between syntax-based tree structures that have to meet highly specific conditions specified in a large number of formation rules. The lexicalist approach requires instead that a small number of rules builds structures by unifying features contained in the lexicon. This trend is found in LMT, UNITRAN, and in general-purpose NLP systems (CLE, PLNLP, ELU) that are being used increasingly for MT. As a consequence, constructing dictionaries from a variety of sources and for different NLP applications has become an important focus for current research (an example is the Japanese EDR project). Further, the interest in language corpora has stimulated connectionist approaches, in which programs 'learn' to recognise links between syntactic patterns and lexical items in semantic networks. It is likely that future systems will be hybrids of linguistic rule-based, statistics-based, and example-based methods.

Most recent developments include the increase of user-designed commercial MT in specific subject domains, and research into spoken language translations systems (notably the C-STAR consortium and VERBMOBIL). Overall, it has emerged that MT research has become more global, with several projects being undertaken in Asian countries. MT is also being more widely used by industry and government agencies. Overall, MT is moving away from being a stand-alone application: it is now being integrated within the professional translation environment ('the translator's workstation') alongside terminology look-up and wordprocessing systems.

Current research and developments

After noting the role of unification grammars in MT research, Jörg Schütz (Saarbrücken), reported on a project to control and improve MT by making terminological information available to a German-English translation system in a concise and efficient form. The approach, which is modular and language-independent, has been implemented in ALEP, an

NLP development platform supported by the LRE programme of the EU. Information in the terminological database is made available to the parsing components in order to resolve ambiguities and to build a semantic representation of the input sentence. The semantic representation in turn forms part of the input to the target language synthesis component.

Dieter Huber (Mainz) supplied a summary of his proposals for using prosodic information in MT. The aim is to draw up a unified scheme for describing and classifying prosodic phenomena which may be applied within a limited English-Japanese translation system.

Horacio Saggion and Ariadne Carvalho (Brazil) reported on their work on automatically translating scientific abstracts in Portuguese. Using a transfer and interlingual approach, they propose concentrating on the resolution of anaphoric references, since these comprise the principal bottlenecks in most systems.

Richard Morgan, Mark Smith, and Sengan Short (Durham) reported on LOLITA, a large, domain-independent and general-purpose natural language engineering (NLE) core developed over the last eight years. LOLITA is being used to translate from Italian into English. The NLE approach is less concerned with implementing theories of computational linguistics than with using a wide range of techniques to build a working system. From a source text input, the prototype system extracts the content (which is represented as a normalised conceptual graph) and the style (expressed as a set of parameters). These are then reconstructed in the target language without further reference to the original. Core components of the system are the semantic network (representing the concepts known to the system), the parser (for morphological and syntactic analysis), the semantic analysis (representation of the text content in a semantic net), and the generator (which converts the nodes of the semantic net into target language expressions).

Walther von Hahn and Galia Angelova (Hamburg) reported on the development of a German-Bulgarian knowledge-based MAT system. The domain knowledge is represented independently of a particular language by a set of conceptual graphs. The user accesses this knowledge by a menu system which establishes links between the knowledge base (KB) and the translation equivalents contained in a separate lexicon. Where direct translations cannot be established, the system returns an equivalent phrase or interpretation that is consistent with the domain knowledge. The advantages of the system are twofold. First, the KB can be updated and edited. Second, the translator is able to clarify the results by querying the generated answer, thus exploring the concepts in the KB itself in order to arrive at the most suitable equivalent. Implemented domains are: banking, motor mechanics, legal arbitration, and chemical devices for oil separation.

Wilhelm Weisweber (Technische Universität Berlin) described KIT-FAST, an experimental German-English MT system which operates with various levels of representation. The system works with four programs: the first is for morphological analysis and synthesis, the second a GSPG parser for syntactic analysis, semantic and conceptual analysis; the third is a transfer and generation module; the fourth is for the evaluation of anaphoric relations. Semantic structure is represented by functor-argument structures and conceptual structures for sentences by the ABox Tell language. There is a separate component for representing text and background knowledge, which is used for anaphoric resolution. The system evaluates pronoun-reference according to the factors of proximity, binding, themehood, parallelism, and conceptual dependency: every possible antecedent is evaluated and a preference score assigned which varies with text type. The treatment of pronouns and the representation of textual knowledge are seen as affording promising perspectives for future development.

Anthony McEnery, Michael Oakes and Roger Garside (Lancaster) supplied a paper on the CRATER project (CRATER = Corpus Resources and Terminology Extraction). A major goal of the project is the automatic construction of bilingual lexica by aligning text corpora and extracting lexical cognates; that is, pairs of words which are reliable translations of each other. Corpora alignment involves establishing which segments of text correspond to each other in a bilingual corpus. The Lancaster research show that the existing language-independent algorithms for this task can be greatly improved by including language-specific information for a particular language pair. This is achieved by using approximate string matching techniques in order to determine the structural similarity of words in different languages. From this, cognates can be established and used as anchor points in the corpora.

Iris Höser and Barbara Rüdiger (GMS, Berlin) described progress on developing a Russian-German MT system based on METAL. A prototype running on a Sun Sparc station is available, and it is hoped to launch a commercial version in two years. Incorporating Russian into the existing METAL software has presented a series of challenges. Apart from integrating the Cyrillic character set (for which METAL was not originally designed), problems arose from Russian's extensive morphological system and from various syntactic features unknown to Germanic languages, such as zero copula in the present tense, subjectless main clauses, interrelation of tense and aspect, missing articles, and complex noun phrases. For analysis, METAL uses an augmented phrase structure model to build X-bar trees with labelled nodes for each phrase and clause. Structural changes are made during the transfer phase. The developers plan to integrate eventually other Slavonic languages into METAL.

Bärbel Ripplinger (Saarbrücken) outlined a proposed architecture for a knowledge-based MT-system as exemplified in VERBMOBIL, a long-term project on the automatic translation of spoken dialogues. The aim is to produce a portable speech-based interpreting device which translates on demand unknown words or phrases into English, of which it is assumed the speakers have at least a passive knowledge. The subject domain is negotiation dialogues for appointment scheduling. The language directions are German-English and Japanese-German. Unlike other speech MT systems under development (SLT, NADINE, ASURA), VERBMOBIL's architecture aims to incorporate pragmatic linguistic information and to operate on semantic representations that are not language-dependent, thereby reducing the number of explicit transfer rules between languages. These representations are derived from the system's KB, or 'domain model'. Organised as a concept hierarchy, the KB contains as much information as possible that is common to both languages (which is what, in the developers' view, makes it language-independent). Information that is specific to the language-pair (such as the translation of the German preposition 'nach' by either 'to' or 'after' in English) is handled by the transfer component. The semantic representation of an input utterance includes pragmatic information, such as whether the utterance is an assertion or a proposal, and its level of politeness.

Scott McGlashan (Saarbrücken) considered the general design principles of spoken dialogue systems, such as SRI's Spoken Translator System, and SUNDIAL. Recognition of spontaneous speech is currently feasible only with vocabularies of about 15,000 words, and accuracy still falls well short of 100%. However, since users are prepared to make allowances when interacting with a computer, it is possible to construct limited but useful systems that are domain-dependent and task-orientated. Typically, a semantic analysis component identifies the domain objects in the dialogue and establishes changes in the conceptual relationships between the objects as the dialogue progresses. A pragmatic analysis component determines the illocutionary value of utterances: that is, whether they are requests, confirmations, etc. Since domain-specific information can be clearly separated from language-dependent operations, it is possible to develop generic components for different languages. Translation is seen as a combination of the interlingual and transfer approaches: the domain-specific semantic representations (the interlingua) are transformed into language-dependent conceptual structures by transfer rules. The relationships between the components are being currently explored in the VERBMOBIL project.

In VERBMOBIL, a speech recognition component constructs a word lattice from voiced input. On the basis of this lattice, a parser builds a sequence of well-formed syntactic structures, from which semantic representations are constructed and utterance types assigned. In order to guide recognition, the system predicts which utterance type may follow at a particular point. Since VERBMOBIL is a dialogue mediator, not a dialogue partner, it must be able to operate without access to the full context of the dialogue. It must also take account of the fact that its users may not be native English speakers. VERBMOBIL is an advance on previous dialogue systems in two respects. First, it employs a theoretical semantic model that is better able to handle discourse structure (that is, beyond sentence level). Second, its provides mechanisms for representing interpretations which depend on and change with context.

Continuing the focus on VERBMOBIL, Susanne Heizmann (Hildesheim) reviewed the characteristics of dialogue interaction and the strategies adopted by the human interpreter. Aspects of these processes can be modelled successfully by an interpreting machine: examples include communicative goals, stereotypical dialogue situations, probable sequences of speech events, and, of course, grammatical information. Other information, such as the ways in which a dialogue dynamically responds to non-verbal behaviour, cannot be modelled. The VERBMOBIL prototype can handle dates, temporal expressions, and speech event types. It is also be programmed to respond with a fixed level of politeness.

Ruslan Mitkov (Saarbrücken) outlined various projects being undertaken by the IAI in Saarbrücken. One of these is CAT2, which originated in EUROTRA. Both an MT system and a software platform for developing grammars, lexicons, and translation modules, CAT2 has been used experimentally for the linguistic analysis and translation of most European languages. The system employs a unification formalism to build and transduce tree structures. Recently, an industrial German-English MT system has been built with an enlarged German morphological component and a dictionary of 3,000 entries in the domain of data-processing. CAT2 is also being used in ANTHEM, a project to produce a multilingual environment for medical diagnoses. Like many MT systems, CAT2 experiences difficulties with anaphora resolution. Plans for resolving the problem include restricting the system to sublanguages (domains) and integrating various sources of linguistic information (ranging from syntactic and semantic data to heuristic and discourse knowledge). The developers have also studied approaches to automatic translation at paragraph level. They propose analysing the source paragraph as a schema of 'rhetorical predicates' and generating the target paragraph as a different set of predicates. The predicates will specify various linguistic functions, such as whether the predicate identifies an entity or amplifies on information already given. The functions will also vary according to the subject domain being analysed.

Christian Boitet (GETA, Grenoble) reported on LIDIA, an Interactive Dialogue-Based MT (DBMT) system implemented on personal computers and designed to be used by non-translators. The idea is that text is sent to an analyser, which asks the author to resolve ambiguities in the source language that will affect the quality of the eventual translation. The ambiguity-free text is then passed on to transfer and generator programs for automatic translation into high quality output. The prototype has been developed for French into German, English, and Russian. The system can be used in a distributed processing or e-mail environment, with the analysis and translation components operating remotely from the user. In this way a text may even be translated into a third or fourth language in succession. Any fresh ambiguities that may be introduced into the target language as a result of translation are resolved by reference to the concept of the 'self-explaining document'. The reader simply runs the analyser for the target language, which runs a similar disambiguation dialogue as for the source language. Although highly abstract interlingual structures are used for the deepest level of conceptual representation, less abstract representations are more suitable for interactive disambiguation, since they more closely parallel source or target language structures. Full-scale multilingual DBMT systems would require very large grammatical and lexical knowledge databases. At the same time, monolingual analysers and 'text-explainers' could be developed on a groupware basis. They could also have a variety of authoring applications apart from MT.

Operational systems

Michael Blekhman (Kharkov, Ukraine) supplied an update on the Russian-English PARS MT system. The first commercial version of PARS was marketed in 1988. The bidirectional PARS-2 system for PCs appeared in 1991 and is widely used throughout the Ukraine and the former USSR. PARS-3 for MS-DOS, Windows and networks was developed in 1994. The latest version has aimed to achieve maximum user-friendliness and has incorporated Borland-type interfaces. Help menus in both Russian and English have been expanded, and a flexible editor included that is compatible with standard Windows wordprocessors. Dictionary management has been improved in two ways. First, the bidirectional dictionaries are fully convertible: entering an item in one direction automatically sets it for the other. Second, morphological and syntactic recognition routines allow full initial entries to be lemmatised and automatically encoded into word type, stem, and other grammatical information. The system comes with a 200,000 bidirectional dictionary. Future systems will include a 25,000 word Russian polytechnical dictionary, and additional subject dictionaries and language pairs are projected. PARS-3 combines interlingual and transfer approaches, and makes use of limited semantic categories. Future development will see a much more powerful transfer grammar and example-based semantic disambiguation.

Angeliki Petrits reported on the current status of SYSTRAN in the European Commission, where it has been used since 1976. The system handles seventeen language pairs, but is employed mainly for English and French (the core languages of the system), followed by German and Spanish. SYSTRAN is used mainly for (a) the fast translation of short, standardised texts, (b) browsing through texts written in a language unfamiliar to the user, and (c) drafting, in which the originator of a text uses the system to produce a draft translation in another language. The system is made up of programs written in assembler and in SPL (the Systran Programming Language) and of dictionaries (basic one-word dictionaries and contextual dictionaries for translating words or expressions according to context). Enriching SYSTRAN's dictionaries with entries from the EURODICAUTOM database has greatly increased the lexical coverage. The system also has access to the EU's legal database, CELEX. An evaluation in 1991 concluded that SYSTRAN should remain the preferred MT system for the Commission and be enhanced with new language pairs. Furthermore, it should be incorporated into an integrated document handling environment and be supported by a post-editing service.

Terence Lewis (Hook & Hatton Ltd) described the Dutch-English system developed by his own company to translate technical documents in the chemical industry. Regarded by its developers as empirical and as embracing no particular linguistic theory, the system uses a large number of grammatical and semantic rules in conjunction with a dictionary database. The processing sequence includes: pre-editing and word look-up; phrase and idiom matching; pattern recognition (for example: bringing together discontinuous constituents of a phrase); specialist dictionary look-up; general dictionary look-up; rule application (for grammatical operations, word order resolution, and disambiguation); and problem-solving routines. The components are modifiable and provide usable, inexpensive MT output with low-cost computer power.

Chadia Moghrabi (Moncton, Canada) described an implemented system for translating cooking recipes between French and Arabic. The system has been under development since 1979. From an input sentence an analyser produces an interlingual semantic network of conceptual structures: concepts, which are stored in dictionaries, include actions (for example: 'put'), elements ('milk') , and qualifiers ('hot'). ATN grammars construct syntactic graphs, and there are morphology-handling modules for French and Arabic. Using stylistic rules, the generator module can produce a number of translations from the same conceptual structure (thus: 'heat the milk, then add sugar to it'; 'add the sugar to hot milk'; 'dissolve the sugar in hot milk').

Svetlana Sokolova (St Petersburg) demonstrated the English-Russian MT package STYLUS for PC (DOS and Windows). The system eschews the conventional stages of analysis, transfer and generation. Translation proceeds instead layer by layer in a hierarchical, 'object-oriented' fashion; that is, from word, to word-group, to clause, and to sentence level. Since the units of each layer are translated without reference to each other, the translation process is controlled and stable: if a unit in one layer cannot be translated, the previous layer's results are used instead. Finite state automata are used to translate the words layer, ATNs for processing the group layer, and procedural frame tools (recognising categories such as verb, subject, object, and complement) for the clause layer. The bilingual lexicon can be updated by the user and includes a utility for automatically generating stems and morphological classes (the maximum size is 25,000 entries). The suppliers claim translation speeds of 150 words per minute, with over 80% high quality output. Interactive, batch and background processing are supported, and the package can be integrated with standard wordprocessors. Over 5000 users are claimed.

The user's perspective

Discussing the reasons why large-scale MT systems (SYSTRAN, LOGOS, METAL) are so little used, Ursula Bernhard (Germany) points out that they remain expensive, are tied to specific hardware, and require specialised personnel. They are hard to use, and it is a difficult and time-consuming task to integrate them into the working environment. Finally, their usage is restricted to technical texts which have a homogeneous terminology. Translators, whose training is not generally technology-orientated, still regard them as job-killers and are rarely consulted about their introduction into a company. They receive inadequate training in their use, and, like the end-users, are often disappointed with the quality of the results. The situation could be helped by improving interfaces with other software and integrating systems into working environments. Dictionaries should be easier to update, and systems should be more robust (requiring less pre- and post-editing of texts). Decision-makers must become more realistic and be prepared to develop innovative applications for MT (browsing, drafting, E-mail, etc.). For their part, developers should respond positively to users' feed-back in order to improve their systems.

Evaluation

Drawing attention to the importance of evaluation in MT, Lorna Balkan (University of Essex) described the EC-funded EAGLES and TSNLP projects which were set up to provide a suite of tests for NLP applications. Current evaluations are based on text corpora (where the language data has occurred 'naturally'), on test suites (the inputs to the evaluation program are artificially constructed), or on test collections (certain outputs are expected from particular inputs). Evaluation may be diagnostic (it aims to localise deficiencies in an application), progress-orientated (the developer wishes to compare the developmental stages of a system), or adequacy-orientated (does the machine meet pre-specified requirements?). Test suites may be constructed bottom-up (the starting-point is the functions of the system itself), top-down (a set of linguistic phenomena is constructed without initial reference to the computer application), or a mixture of both. The TSNLP project aims to establish guidelines for constructing test suites, to produce test suite fragments covering core syntactic phenomena in English, French, and German, and to develop tools for building and using test suites. TSNLP will adopt a top-down approach and include an annotation scheme which will provide precise information about the input and output (such as length of sentence, syntactic category, position of constituents, and type of error).

Proposals

Boh Wasyliw (De Montfort University) and Douglas Clarke (Cranfield University) presented a number of cases in which misunderstandings arising out of verbal communications between pilots and air traffic control (ATC) have led to serious accidents. An ideal solution would be a fully automatic computer-based analysis engine functioning as the communications interface between pilot and ATC. Such an engine would comprise a voice input and speech recognition module, an analytic module (which would filter the spoken input and identify ambiguities), and a visual output module (this would present the alternative interpretations to the pilot for selection). Even a monolingual system of this nature is, of course, well beyond the bounds of current applications. It would require complex software performing sophisticated linguistic analysis at the phonetic, syntactic, semantic, and pragmatic levels. Processing would also have to take place in real time. A possible system could eventually incorporate translation modules, enabling the pilot and the controller to think and speak in his native tongue. As an interim solution, communication between pilot and ATC could be conducted via screens of strictly controlled information presented visually in the form of menus of possible and unambiguously understood actions.

In conclusion, Alan Melby (Brigham Young University) outlined his views on the type of software required for the long-term development of high quality MT. In Melby’s view, such software would allow for the fundamental ambiguity, flexibility and dynamic metaphor inherent in natural language, and be ultimately non-algorithmic in character.

CAT2 — A UNIFICATION-BASED MACHINE TRANSLATION SYSTEM

by

Ruslan Mitkov

CAT2 is a MT system embodying a unification-based formalism, similar to PATR-II (Sharp 1988, Sharp 1991) and software for the development of grammars, lexicons and translation modules. It was developed at the IAI, Saarbrücken, as a sideline implementation to the Eurotra Project, and has been undergoing constant development and evolution since 1987. Experimental versions of numerous languages have been implemented, including English, German, Spanish, French, Portuguese, Italian, Dutch, Russian, Greek, Korean, and Japanese. It is now being used in pre-industrial projects for a number of commercial firms and academic institutions (Sharp and Streiter 1995).

The translation strategy is based on tree-to-tree transduction, where an initial syntactico-semantic tree is parsed, then transduced to an abstract representation ('interface structure') that is designed for simple transfer to a target language interface structure. This structure is then transduced to a syntactico-semantic tree in the target language, whose yield provides the actual translated text. The analysis of a source language, as well as the generation of a target language, is based on strictly monolingual rules, the transfer component being the only interface between two languages. Thus, an analysis in one language may be transferred to any number of target languages without requiring re-analysis. The various components, however, may make use of common rules, much like subroutines, so that 'universal' descriptions may be made to apply equally to any number of languages, thereby reducing the rule base tremendously, as well as simplifying the maintenance of grammars and the addition of new language components.

The formalism specifies two rule types for tree construction, and two rule types for tree transduction. Trees are built using 'b-rules', a context-free backbone with attribute-value pairs rather than simple category symbols. The following illustrates how the rule 'S → NP VP' might be written in CAT2 notation:

```
(a)    {cat=s} .[ {cat=np}, {cat=vp} ].
```

The feature bundles may include any number of simple or complex feature descriptions; simple features have atomic values, for example: 'cat=s', whereas complex features have feature bundles as values, for example: 'agr={num=sing,per=3}'. In addition, since the formalism is implemented in Prolog, a value may be a logical variable, bound to another variable with the same name within the rule; instantiation of one of the variables automatically instantiates the other. The implementation also allows for negative and disjunctive features, implemented in SICStus Prolog using the 'when/2' construct for freezing goal evaluations.

The second rule type in tree construction is the 'f-rule' for validating the feature content of partial trees. A simple f-rule for ensuring subject-verb agreement might be coded as follows:

(b) {} .[{cat=np}>>{agr=X}, {cat=vp}>>{agr=X}].

This rule states that, in a tree configuration containing an NP as left daughter and a VP as right daughter, their agreement features must unify.

In practice, our grammars make use of a very small number of b-rules, based on X-bar syntax, (extended) head features (Streiter 1994), and lexically-driven tree construction. The f-rules instantiate various universal and language-specific principles and properties, as well as supplying default values to lexical and phrasal constructions.

The tree transduction rules employ analogous rule types: t-rules transform tree structures, and tf-rules copy or transform selected features from source to target trees. The rule formats are similar to b- and f-rules, and again unification underlies the rule application. Since the rules for anaphora resolution in our model do not employ t- or tf-rules, they will not be further described here (see Sharp 1994 for a complete description of the formalism).

Recently, CAT2 was extended to be able to handle pronominal anaphora (Mitkov, Choi and Sharp 1995).

A pre-industrial prototype of CAT2 has been developed for a dictionary of about 10,000 entries (mainly data processing) in the language pair German-English. For other language pairs (for example, French-German/French-English), experimental versions are available. The prototype accepts free input, especially in the German source pair.

Since 1987 CAT2 has been used in various universities as a teaching device and for the definition and processing of language analysis, synthesis and translation.

Taking over the long tradition of the University in Saarbrücken in 'electronic language research', the IAI is currently carrying out, apart from CAT2, a number of other application-oriented projects sponsored by the LRE and MLAP programmes of the Commission of the European Union, German ministries, and by private industry.

Among these are projects using the most advanced techniques in NLP, such as typed feature structures, for example, ALEP, the Advanced Language Engineering Platform). Other projects are using the well tested CAT2 prototype for industrial validations.

References

Mitkov, R., Choi, S. K., and Sharp, R (1995) 'Anaphora resolution in Machine Translation' (in press)

Sharp R. (1991) 'CAT2: An Experimental Eurotra Alternative', *Machine Translation,* Vol. 6: 215-28.

Sharp, R. (1994) *CAT2 Reference Manual, Version 3.6,* IAI, Saarbrücken

Sharp, R. and Streiter, O. (1995) 'Applications in Multilingual Machine Translation', paper submitted to *Practical Applications of Prolog*, Paris.

Streiter, O. (1994) 'Komplexe Disjunktion und erweiterter Kopf: Ein Kontrollmechanismus für die MÜ', *Proceedings of Konvens '94*: 28-30, Vienna

PRACTICAL ASPECTS OF THE USE OF METAL AT SIEMENS NIXDORF

by

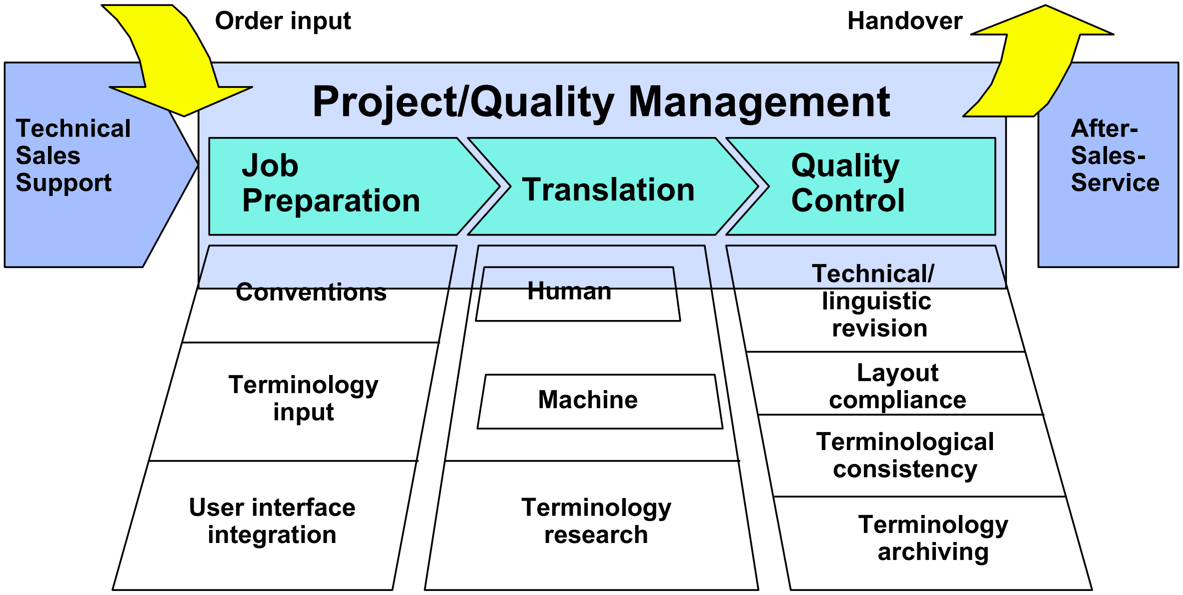
Keith Roberts

The following account is based on a lecture given by Keith Roberts, Manager of Language Services, Siemens Group Services Limited, Chertsey, at King's College London at the AGM of the Natural Language Translation Specialist Group of the BCS on 22 September 1994.

The METAL machine translation system originated at the Linguistics Research Center in Texas, USA in 1959-60, when work on a German-English system was initiated. The German company Siemens took over the project in 1979 when it entered the American market for analog switching equipment in the telecommunications industry. LRC's involvement ceased in 1994, and the system is now under the control of Siemens Nixdorf (Sietec), who continue to use it within the framework of their Munich- and Paderborn-based translation service, the *Sprachendienst*.

The *Sprachendienst* employs about 80 people, has around 2,200 customers, and translates approximately thirty-four million words a year in the fields of information technology and telecommunications. The service is organised as a distributed translation environment, the local production centres having remote access to METAL and local access to the TWIN terminology management system.

Some time was devoted to describing the key project stages involved in translation at Siemens Nixdorf and to the integrated translation process (or 'process chain' as it is often referred to) at the *Sprachendienst*.



Particular emphasis was placed on how Machine Translation can be integrated into the documentation process:

- MT as a service
- Various levels of post-editing in line with customer requirements
- Terminology projects for subsequent MT usage
- Project solutions
- Consultancy, including training

Written in LISP, METAL has been ported to a SUN platform and is capable of processing up to 1,000 pages a day in machine-time. The METAL Database Management System comprises four modules: the source language (SL) dictionary, the transfer dictionary, the target language (TL) dictionary, and the grammar rules. The actual number of grammar or linguistic processing rules is relatively small (about 550).

The typical sequence of operations for throughputting text in METAL was described. The first phase is text acquisition and preparation, which involves deformatting the source text, i.e. removing features of layout and reducing the text to translation units comprising its constituent words. The second phase is lexical analysis, or dictionary look-up; any unknown terms are entered manually or semi-automatically into the lexicon. Translation is performed during the third or machine run phase. The fourth and final phase consists of post-editing and reformatting.

For the linguistic representation of the source language, METAL employs a dependency tree model. The system generates a parse tree for the SL sentence, which is then converted into a representation of the TL sentence by the transfer component. To translate compound noun phrases, a terminology look-up component searches for the longest string in the SL for which there is also a TL equivalent.

No presentation on the METAL Machine Translation System can dispense with at least a broad outline of the theory of MT, but thereafter the talk concentrated strictly on practical experience gained within the *Sprachendienst* over a number of years:

a) Where MT works best

In SNI's experience specific subject areas and language pairs are required, involving high-volume texts and, wherever possible, the use of restricted language, all of which should be coupled with a continuous assessment of output.

b) Suitability for MT

Emphasis was placed on the need to assess the source text for MT friendliness in terms of language pairing, document style, volume, subject area etc. Concrete examples of both good and poor quality MT translations were presented and discussed.

c) Typical problems

Examples were given of some of the difficulties encountered at the stages of pre-editing (for example: terminology not coded adequately, non-translatable units not masked) and post-editing (for example: level of post-editing needed, text editor used, experience and attitude of human post-editor).

d) Potential solutions

Recent efforts by the *Sprachendienst* to introduce the use of controlled language among its key accounts were described in some detail. 'Writing with the translation process in mind' concentrates on both human translation (at sentence and word level) and MT (for example: no combinations of words and special characters such as <$Date>). While a certain amount of success has been obtained to date, it is at times proving difficult to quantify the benefits, and technical authors and developers tend to perceive this approach as a restriction on their creativity.

How SNI's *Sprachendienst* has attempted to introduce MT into a practical translation environment was set out in a detailed description of an in-house development known as DTS (Distributed Translation and Terminology Services). DTS can be accessed quickly and simply by translators, suppliers and customers via modem and LAN and offers a whole range of services, such as

- Access to MT (METAL)
- Access to terminology services (glossary compilation, dictionary look-up and updating etc.)
- Automatic volume count
- Automated assessment of MT friendliness
- Link to invoicing system

The final part of the talk was devoted to the issue of Project Costing, demonstrating how costings are built up at SNI with a view to ensuring — in certain cases substantial — cost savings for the customer through the appropriate use of MT/DTS.

In conclusion, it was clear that METAL is being used commercially within the Siemens organisation. It is regarded as one of a set of tools which may contribute to improving operational efficiency; as such its output is subject to continuous assessment and evaluation.

LINGUISTIC RESOURCES ON THE INTERNET

by

Roger Harris

General linguistic resources, and machine translation resources in particular, may be found in many parts of the Internet. Similar networks, such as Compuserve, also have linguistic sections (Compuserve: GO FLEFO or GO MACCIMSUP, which uses translation technology — Eng/Fre/Ger — from Intergraph Corp.), but these are not directly available on Internet.

Access to much of Internet's data may be obtained with a simple computer, modem and communications program. My own ‘antique’ system includes an Amstrad 1640 connected to a 1200 baud modem. Where linguistic resources include high-quality screen graphics, colour, and sound, then obviously you will need a more advanced computer system to take advantage of these features.

You may gain access to Internet either via a terminal connected to a large academic or corporate computer system or via a stand-alone computer connected to the telephone system. Whatever the case, the distant host computer to which you will be connected will have a directory structure similar in some ways to PC-DOS as used in IBM-PC's and their clones. If you are accustomed only to Windows or Apple Macintosh icon screens then you might find this a problem. However, navigation is very simple and only a few commands will be needed.

Your computer will be connected to your local host computer and will exchange data using a protocol such as Kermit, XMODEM or YMODEM. The distant host computer, for example, in Brazil, will automatically set up a communications link with your local host. A little knowledge of the Unix operating system might be useful but is not essential. Other than that, instead of accessing, say, drive C: or drive F: on your computer, you will be accessing a distant hard disk drive which is identified by name instead of by letter. It's also a bit slower. That's all.

Various computer commands are shown below enclosed in single quotation marks, for example, 'dir'. The quotation marks are not to be used when typing in a command. The control key is represented by ^, as in '^C' for control C. The Carriage Return or Enter key is represented by <CR>.

Help in the form of screens of commands and explanations is available almost everywhere by typing 'h' or 'help' or '?' at the prompt and pressing <CR>; use data logging (see below) to build up your own file of help pages. Type 'cd' pathname <CR> to select a directory path. Type 'cd ..' <CR> to go up one directory level. Type 'cd' to go to the highest directory level. Type 'dir' <CR> to display the current directory which will probably look something like this:

```
MULBERRY.SRV.CS.CMU.EDU:/usr0/anon
drwxr-xr-x   2 root   system     512  Aug  24  1994  misc
-rw-r--r--   1 root   system    1158  Nov  23  1993  READ_ME
drwxr-xr-x   2 root   system     512  Feb  17 00:33  project
drwxr-xr-x   6 3973   0         2048  Aug  24  1994  sys
drwxr-xr-x   2 root   system    1536  Feb  21 15:43  user
```

In the line above the body of the directory, the capitalised words refer to the full name of the computer being accessed (CMU = Carnegie Mellon University), while /usr0/anon is the directory path.

In the body of the listing, in column one, if the first character is a ‘d’ (as in, for example, drwxr-xr-x), then the line refers to a directory which may in turn contain other directories and/or files. If the first character is a ‘-’, then the line refers to a file. Letters after the first character refer to file security and access.

In the second line, READ_ME is a text file which can be downloaded into your host computer by typing 'get' READ_ME <CR>. The number 1158 is the number of bytes in that file and Nov 23 1993 is the date when it was last updated. Type 'cd' project <CR> to go to the project directory.

Usually, one can type in the full directory path. If that is rejected then type in the directory path names individually until you reach the desired directory. To examine the contents of a directory, type dir <CR>, to check in case the next directory's name is not the same as that in your reference source.

The distant computer could take up to a minute or more to respond depending upon data traffic levels. Repeated, impatient pressing of a key will significantly slow down the response; each keystroke is stored and will be executed once the distant computer is free to deal with your instructions. It may have to execute the dozens or even hundreds of keystrokes which you might have made. This could take minutes or more and the distant system might appear to have gone berserk. Sometimes it is best to hang up, type '^H' <CR>, and log on again.

Be prepared to use your ingenuity in finding files and do not be surprised if the directory structure which you encounter does not exactly match the description obtained from a reference book or from this article. If you get lost or need help, then type 'h' or 'help' or '?' and press <CR>. Internet is an evolving entity subject to continual change, revision, addition, and deletion.

Data logging

Any text which appears on your screen may be sent to a printer by pressing the <PRT SCRN>key while holding down the <SHIFT> key. The communications software will automatically store in a named file any text which is received from a distant source. This process may be called data logging. The software will store the text in a file whose suffix or extension name might be .LOG, or you can specify a filename. The data logging function must be invoked before the data appears on your computer screen.

A source of unique albeit cryptic filenames is based upon the date and time when you are about to store some data. For example, 7.33 pm on 16th February 1995 may be coded as 5=year (1995), 2=month, 16=day, 19=hour, 33=minutes, to give the filename of 52161933. Use A, B, and C to represent October, November, and December. Add the filename extension '.199' if you want to identify the decade. Such filenames will appear in date/time order in a sorted display of filenames.

Archie, Gopher and Veronica

These are programs which will perform keyword searches in order to locate files and programs; archie operates on ftp sites and veronica on gopher sites. Archie will only locate a source; veronica will locate a source and retrieve a file. A search allows Boolean parameters such as 'machine AND translation', machine OR translation', and 'translation NOT machine'.

An archie search for 'linguist' will return addresses such as:

coombs.anu.edu.au /coombspapers/coombsarchives/linguistics/
csli.stanford.edu /linguistics/
ftp.sunet.se /pub/mac/umich/misc/linguistics/
ftp.univie.ac.at /systems/dos/simtel/linguist/
julian.uwo.ca /doc/FAQ/greek-faq/linguistics
knot.queensu.ca /pub/tcrunchers/Misc/linguist.list

Use the data logging function of your communications software to store the data as it is displayed.

USENET newsgroups

There are reputed to be some 7,000 to 8,000 newsgroups on Internet. Your Internet access provider may be connected to less than half that number. Some newsgroups are either empty or dormant. A newsgroup is composed of groups of comments called threads which in turn are composed of articles put there by posters (writers). You can post a reply directly to an article on the thread or to the poster. You can download articles or whole threads. A file in your computer may be uploaded and then posted to an existing newsgroup thread or to one which you can initiate. The following newsgroups are of linguistic interest:

alt.etext	electronic texts
comp.ai.nat-lang	Natural language processing by computer
comp.ai.nlang-know-rep	Natural language and knowledge representation
comp.software.international	Finding, using, and writing non-English applications
comp.speech	Research and applications in speech recognition and production
comp.text.sgml	Structured documents markup languages
sci.lang	Natural languages, communication
sci.lang.translation	Problems and concerns of translators

FAQs (Frequently Asked Questions) are extensive, detailed documents providing copious information about the subjects covered by a newsgroup. They are available at:

ftp: rtfm.mit.edu /pub/usenet/newsgroup's name

where ‘newsgroup’s name’ is, for example ‘sci.lang’.

Linguistic resource sites

In the following section, file locations are shown as follows:

protocol: computer_name /directory/directory/file

Computer network addresses are often shown as four numbers in addition to the more usual letters, for example:

ftp.cmu.edu [128.2.206.173].

Either format should work. All directory names are case-sensitive. You first log on using the computer name and then select the directory. The following is a list of useful linguistic resource sites:

Alex Catalogue of Electronic Texts

http://www.lib.ncsu.edu stacks/alex-index.html
gopher://rsl.ox.ac.uk 70/11/lib-corn/hunter
gopher://gopher.lib.nmsu.edu /11/library/stacks/Alex

Association for Computational Linguistics (ACL)

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

Send e-mail to

listserv@cs.columbia.edu

with the following in the body of the message:

index acl-1

Brown University linguistics page

http://www.cog.brown.edu /pointers/linguistics.html

Colibri newsletter (language, linguistics, etc.)

http://colibri.let.ruu.nl

comp.ai.nat-lang Usenet newsgroup

Dragomir Radev is the editor of comp.ai.nat-lang FAQ and the source of several items in this list.

e-mail: radev@cs.columbia.edu
http://www.cs.columbia.edu /~radev/home.html

Consortium for Lexical Research

e-mail: lexical@nmsu.edu
ftp: crl.nmsu.edu /CLR/catalog

Corpora, dictionaries, wordlists etc.

e-mail: ingrid.maier@slaviska.uu.se	(Russian corpus)
e-mail: ldc@unagi.cis.upenn.edu	(CELEX, LDC)
ftp: black.ox.ac.uk /wordlists/	(word lists)
ftp: ftp.cmu.edu /project/fgdata/dict/	(dictionaries)
ftp: ftp.cs.vu.nu /dictionaries	(word lists)
ftp: ftp.funet.fi /pub/doc/dictionaries/	(word lists)
ftp: ftp.uu.net /doc/dictionaries/DEC-collection/	(dictionaries)
ftp: ftp.white.toronto.edu /pub/words/sodict.gz	(Shorter Oxford)
ftp: gatekeeper.dec.com /pub/misc/stolfi-wordlists	(word lists)
ftp: wocket.vantage.gte.com /pub/standard_dictionary	(word lists)
http://olymp.fer.uni-lj.si /dictionary/a2s.html	(Eng.-Slovene)
http://philae.sas.upenn.edu /French/french.html	
http://solar.rtd.utk.edu /friends/cyrillic/cyrillic.html	
http://www.fmi.uni-passau.de /htbin/lt/lte	(Eng.-Ger. dictionary)

Echo Eurodicautom

Translates words between Dan/Dut/Fre/Ger/Ita/Por/Spa:

http://www.uni-frankfurt.de /~felix/eurodicautom.html

EITS (Experimental Internet Translation Service)

Launched in 1994, and offered, apparently, translations between many of the world's known languages including Pig Latin and Ubby Dubby. You can get the full hilarious details by e-mailing to jens@panix.com

with the subject line as

request-file eitsfaq.txt.

ELSNET (European Language and Speech Network)

e-mail: elsnet-list@cogsci.ed.ac.uk
http://www.cogsci.ed.ac.uk /elsnet/home.html

International Standards Organisation (ISO), ISO Online

http://www.iso.ch /welcome.html (English version)
http://www.iso.ch /welcomef.html (French version)

ISO-8859-1 FAQ (International Standards Organisation)

ftp: ftp.vlsivie.tuwien.ac.at /pub/8bit/FAQ-ISO-8859-1

Institute for Natural Language Processing at the University of Stuttgart

http://www.ims.uni-stuttgart.de/IMS.html

Lingsoft Corp Inc's demonstrations of their linguistic software

http://www.lingsoft.fi /cgi-pub/engcg	(English parser)
http://www.lingsoft.fi /cgi-pub/engtwol	(English morphology)
http://www.lingsoft.fi /cgi-pub/finhyp9	(Finnish hyphenation)
http://www.lingsoft.fi /cgi-pub/finstems	(Finnish stems)
http://www.lingsoft.fi /cgi-pub/fintwol	(Finnish morphology)
http://www.lingsoft.fi /cgi-pub/gertwol	(German morphology)
http://www.lingsoft.fi /cgi-pub/swetwol	(Swedish morphology)

LINGUIST list

http://www.ling.rochester.edu /linguist/contents.html

Send e-mail to

listserv@tamvm1.tamu.edu

with the following in the body of the message:

‘subscribe linguist forename surname’

You will receive frequent bulletins on various linguistic subjects; a message: ‘unsubscribe linguist forename surname’ will cancel.

Linguistic tools

clarity.princeton.edu /pub/
linc.cis.upenn.edu /pub/xtag/
speech.cse.ogi.edu /pub/tools/

Linguistics and MT document archive and e-print server

http://xxx.lanl.gov /cmp-lg/

Multilingual PC Directory

This book is a copious source of information about linguistics software, wordprocessors, fonts, suppliers’ addresses, Internet resources, etc. (ISBN: 1-873091-03-5). There is also an electronic version in Windows Help File format, available for downloading from Compuserve’s Foreign Language Forum (GO CIS:FLEFO) in the file MPCDIR.ZIP, and also from the site: http://knowledge.co.uk/xxx/, due on-line in May 1995. Contact: Knowledge Computing, 9 Ashdown Drive, Boreham Wood, Herts. WD6 4LZ, Tel: +44 (0)181-953 7722, Fax: +44(0)181-905 1879, E-Mail: 72240.3447@compuserve.com.

Natural language software list

ftp: ftphost.uni-koblenz.de /outgoing/software_list.ps.z

Natural Language Software Registry, Saarbrücken

http://cl-www.dfki.uni-sb.de /cl/registry/draft.html

For a descriptive document and questionnaire:

ftp: crlftp.nmsu.edu /pub/non-lexical/NL_Software_registry
ftp: dri.cornell.edu /pub/Natural_Language_Software_Registry or /pub/NLSR

NL-KR Digest (as published on Usenet newgroup comp.ai.nlang-know-rep)

For subscriptions, send an e-mail request to:

nl-kr-request@ai.sunnyside.com

For submissions, questions etc., send:

nl-kr@ai.sunnyside.com

Back issues are available from the following:

ftp: ai.sunnyside.com /pub/nl-kr/Vnn/Nnn
(where Vnn = volume number, Nnn = issue number)
gopher: ai.sunnyside.com (Port 70) /pub/nl-kr
http://ai.sunnyside.com /pub/nl-kr

Software localisation

ftp:etext.archie.umich.edu/pub/Economics/FutureTalk/media-localising.txt.gz
http://gopher.gmu.edu /bcox/Economics/SoftwareLicensingPaper.html

Translators' Home Companion

http://www.rahul.net /lai/companion.html

Unicode Consortium FAQ

Full listings of ISO 639 and ISO CD 11639, and the use of ISO/IEC 6420 control functions to encode language:

http://www.stonehand.com /unicode.html
http://www.stonehand.com /unicode/standard/principles.html#x12
(deals with language tagging)
e-mail: unicode-inc@hq.metaphor.com

University of Virginia electronic text centre

http://www.lib.virginia.edu /etext.ETC.html

Usenet's 1000 most commonly used words and usage statistics

ftp: ftp.spies.com /Library/Article/Language/top1000.use

This site also contains further files of linguistic interest.

The above information has been collected from Usenet newsgroups, Internet searches, an FAQ edited by Dragomir Radev, and the *Internet Golden Directory*, 2nd edition. My thanks go to them all.

(Roger Harris may be contacted at <rwsh@dircon.co.uk>.)

BOOK REVIEWS

Peter Roach (editor) (1992) *Computing in Linguistics and Phonetics. Introductory Readings*, London: Academic Press. Paperback. £14.95. 115 pages. ISBN 0-12-589340-X

Computing in Linguistics and Phonetics constitutes a comprehensive review of the current computing applications in linguistics. It is a general introductory text aimed at the linguistics undergraduate.

The book is set out in seven sections, each covering a specific area of application. The first chapter begins with an introduction to computing and computing terms. It describes a computer and a set of basic terms.

The second section introduces the application of data banks in linguistics. Their usages are briefly described, with examples of the types of searches that are carried out. The author goes on to describe the structure in which the data is stored and how the structure is related to the retrieval mechanism. The stored text may be spoken or written. Text is tagged with different types of information such as grammatical, phonetic or typographical. Examples of each sort are given.

The third chapter describes the ways in which computers are used to analyse, produce and recognise speech. Speech can be analysed acoustically. Waveforms are represented graphically and are analysed by frequency, cycle and filtering. Applications of this type of work include work improving intelligibility of speech such as that involved in teaching languages or teaching deaf people to speak.

The analysis of speech production is concerned with picking up airflow information which can be converted into an electronic signal that then is processed by a computer. Another area tries to study the changes in dimensions of the vocal tract when a sound is produced by the movements of the jaw, tongue, lips and soft palate. Speech recognition is another major area of research.

The fourth chapter deals with Natural Language Processing. The methods of phrase structure grammar, Transition Networks, Sentence Networks, and Augmented Transition and Sentence Networks are described in some detail.

Chapter five deals with speech production by machine. It describes the process involved in producing a speech signal. Two basic methods are described. These are the Terminal Analog System, which simulates human speech, and the Live Analog System which copies the way in which sound is produced.

Chapter six introduces the subject of Machine Translation. A potted history of the development of MT is briefly discussed, with examples of MT systems to illustrate the points. Translation is described as being a three stage process. These stages are analysis, transfer of the logical representations of language from the source language to the target language. MT and its limitations are also covered.

Computer Assisted Language Learning (CALL) is the subject of the last chapter, which presents a comprehensive survey of currently available software and a concise history of developments in the field. There are also some suggestions for future developments.

Each chapter deals with a field or branch of computational linguistics and as such represents an extensive body of learning in its own right. Since the text is an introductory one, care has been taken to present a broad overview of each area. In some cases this has led to a rather simplistic approach. For example, there are many limitations to the ways in which computers are able to represent or process natural language: computers cannot, for example, translate perfectly between languages, or spontaneously generate sentences in the way in which human beings do. The creative element in language cannot be mirrored by a computer and forms a constraint on all types of applications.

Hardware limitations are briefly discussed in the book, but I suspect that greater emphasis could have been placed on the fact that a computer solution to a language processing problem generally requires a large amount of storage area and power.

To sum up, the book admirably fulfils its main objective, which is to provide an informative guide for the linguistics undergraduate. The articles are all precisely written in straightforward language, and do not require any previous computing experience. The scope of subjects includes applications in most aspects of linguistics, ranging from syntax (phrase-structure grammars), to phonetics and language acquisition. A bibliography of further reading is provided with each chapter. Areas such as historical linguistics and sociolinguistics pose particular problems which are harder to solve by computer, and these are left untouched by the book. Overall, its simplistic approach and breadth of topics make it a highly suitable text for non-computational undergraduates.

Tania Reynolds

Arnold, D., Balkan, L., Lee Humphreys, R., Meijer, S., and Sadler, L. (1994) *Machine Translation. An Introductory Guide*, Oxford: NCC/Blackwell. Hardback £40.00. ISBN 1-85554-246-3. Paperback £18.99. 240 pages. ISBN 1-85554-217-X.

The authors are, or have been, members of the Computational Linguistics and Machine Translation Group at the University of Essex. They have aimed in this volume to provide a genuine introduction into MT theory and practice by affording a broad range of insights into the field: how MT systems work; their capabilities and limitations; methods for evaluating systems; and likely developments in the short to medium future. Targeted at students, translators and managers, the book presupposes no specialist knowledge of computational linguistics. It is clearly organised, with each chapter providing a summary of points covered and recommendations for further reading. A short glossary of technical terms is included, together with a list of useful addresses.

The practical emphasis is evident from the start. The first chapter tackles head-on popular (and occasionally amusing) conceptions and misconceptions about MT. The second follows the passage of a document within a fictional multinational organisation in order to illustrate how MT could actually contribute to the process of translation within a commercial environment. Such perspectives are particularly useful for language students who are contemplating translation as a career but have little conception of either the type of language which is most often translated or of how documents are prepared and handled within large organisations. The status and nature of electronic documents, the role of controlled languages, and the complexities of evaluating MT systems are further issues considered later in the volume which retain the focus on MT as a practical activity.

Perhaps even more interesting for a computationally non-expert academic readership are the chapters describing the architectures and linguistic models underlying MT systems. For those unschooled in phrase structure theory, the outline of the principles of constituent structure and tree diagrams are especially useful. There are clear expositions of basic parsing techniques and of how linguistic information may be represented in either surface-based forms or 'deeper', more abstract and semantics-orientated ways for subsequent processing by the transfer and generation components of an MT program. Such representations are at the heart of current research and development within the wider field of NLP. Individual chapters are devoted to word formation and morphological processing, the structure of dictionaries and lexical information, and to the identification and handling of terminology within an MT context.

Current and future developments are addressed throughout the book. They include the long-standing preoccupation of builders of MT-systems with such problems as ambiguity, anaphoric references, lexical and structural mismatches between languages, and the extent to which real-world knowledge may be represented within the machine. A separate chapter is devoted specifically to new directions in MT. One such direction is the extension of rule-based approaches to incorporate various levels and types of linguistic information. Further directions include the interest in knowledge-based MT, with its focus on subject domains, and the use of existing language resources for MT purposes, in particular corpora and parallel multilingual texts. There is also a useful exposition of how statistical methods are being employed to perform certain MT tasks, such as the disambiguation of word senses.

A most welcome feature of the book is the way in which the authors have taken care to illustrate translation processes with examples from actual languages (my only quibble in this respect is that the German *zurück* is curiously identified at one point as a preposition). This explanatory approach lends the book transparency and gives the reader a genuine feeling as to what goes on under the bonnet of an MT system. While many of the main research and operational systems are mentioned, they are not described in any detail. In view of the extent of the MT field, however, this is wholly in keeping with the character of an introductory guide. In summary, this volume admirably achieves its aims and succeeds in combining breadth of coverage with both theoretical and practical depth. It avoids the mystification so often found in computational linguistics and should indeed stimulate its readers to take a long-term interest in MT applications.

Derek Lewis

Karen Jensen, George E. Heidorn, and Stephen D. Richardson (1993) *Natural Language Processing: the PLNLP Approach*, Boston: Kluwer Academic Press. Hardback. £80. 324 pages. ISBN 0-7923-9279-5.

PLNLP (Programming Language for Natural Language Processing) evolved during the 1970s and 1980s as a general purpose software tool for writing natural language applications. Typically, the developer of the application, a linguist, builds a grammar and lexicon from PLNP rules and data structures in order to parse a sentence. Possible applications range from checking grammar and analysing style to machine translation. Although most applications operate at sentence level, tentative progress has been made in building a model of discourse by linking sentences within paragraphs. The system has various components organised as a continuum and is claimed to provide a highly flexible and integrated set of tools for processing and analysing languages (which include Japanese and Arabic, although most applications are based on English). This volume presents the first published collection of papers (twenty two in all) about the PLNLP system and its applications. Many of the authors were originally members of the group at IBM which pioneered this work at the Thomas J. Watson Research Centre.

NLP systems vary as to their relative reliance on dictionary information or processing rules. Although PLNLP is not committed either way, it is clear that dictionaries are indispensable for many of its components. Of particular interest is the emphasis on employing on-line machine-readable versions of standard lexica from which the computer program extracts the information it needs in order to produce a correct parse or otherwise represent the structure of an input sentence. Most of the papers describe a stage in the processing and its relevance to a particular application.

The component which supplies the initial syntactic sketch of an input sentence, the transductive grammar, aims to produce a parse for any input; that is, it responds to real-world textual data more robustly than parsers in the tradition of generative grammar. The transductive grammar written for English (PEG), operates with a large but unsophisticated lexicon based on Webster's Seventh New Collegiate Dictionary. In the next stage of processing, semantic information is applied to correct, or re-assign, the syntactic sketch produced by PEG. Re-assignment relies on large amounts of information per word held in complete on-line dictionaries, so the lexicon plays an unusually important role for an NLP system. Dictionaries mentioned in the applications described here are Webster's Seventh and the Longman Dictionary of Contemporary English (LODCE). Lexica play no role in the third stage (PEGASUS), where underlying logical forms of sentences are constructed solely from the parse structures derived so far. During the next stage, sense disambiguation , on-line dictionaries are used extensively to narrow down senses of verbs and nouns in the parse. The disambiguation program, called MAST (MApping Senses to Text), works for English by comparing the structure and context of a word in the input text with information in a dictionary entry. Lexical entries in the LODCE follow a strict argument structure which enables MAST to extract the semantic features on subjects and objects of verbs. In this way MAST is able to assign the input word its most probable sense. The final two stages, in which rules and dictionaries interact closely, are still under development. The first of these aims to produce common semantic representations for sentences which mean the same thing (a process started by PEGASUS). This is achieved by a concept grammar which establishes underlying semantic graphs for different surface structures (for example: A is under B is equivalent to B is over A). Lastly, the system identifies links between words and concepts to build a conceptual model of the topic of a paragraph (this is perhaps the most theoretical of the papers).

The flagship application of PEG is the Critique system, which identifies grammar and style problems in English texts. Relying primarily on syntactic rules to parse input and flag possible errors, it looks for the best structural ‘fit’ by progressively relaxing grammatical constraints, and, if no complete parse is found, combining the available analysed sub-structures. The user may tune the system to suit his own profile and be presented with message, help and tutorial information. Critique worked best with standardised non-literary texts containing short sentences and has office, publishing and educational applications.

PEG has been used as a front end for a number of prototype machine translation systems, SHALT (English to Japanese), C-SHALT (English to Chinese) and PORTUGA (English to Portuguese, Norwegian and other languages). A transfer system relying on contrastive knowledge of a language pair, PORTUGA aims for broad-based coverage, i.e. it will handle ill-formed and incorrectly parsed input and is not restricted to a limited domain. Its bilingual dictionary is notable for its relative simplicity (possible translations of a source language word are separately annotated with the conditions for its translation) and multi-word expressions can be imported directly from published dictionaries.

The PLNLP group provides convincing evidence of the potential of machine readable dictionaries as sources of extractable information for NLP applications. The implication is that natural language may be a convenient and rich medium of knowledge representation, for instance, in the form of semantic networks, if we can develop the computational techniques to retrieve the information. Illustrations of this approach, which basically uses the dictionary as a source for setting up a knowledge base, include analysing the semantic structure of head nouns as obtained from dictionary definitions (chapter 10), finding the most likely attachment for English prepositional phrases by comparing them with similar constructions in a lexicon (chapter 11), uncovering the semantic relations between members of compound noun sequences, such as ‘vegetable market’ or ‘lemon peel’ (chapter 13)), and identifying patterns of semantic relations between verbs and their noun phrases as typically found in dictionary entries (chapter 15).

Although the PLNLP group makes no strong claim for a particular model of linguistic representation, it is evident that they eschew a strict rule-based approach to natural language processing and see the most exciting prospects in the extraction and application of lexical information. Although expensive, the volume represents a unique and clearly presented collection of valuable papers.

Derek Lewis

Conferences and Workshops

The following is a list of forthcoming or recent conferences and workshops. Telephone numbers and E-Mail addresses are given where known.

19-22 April 1995
PACLING ‘95, Pacific Association for Computational Linguistics
Brisbane, Australia
Tel: 61 7 365 6896, E-mail: sussex@lingua.cltr.uq.oz.au

28-29 April 1995
1st Annual Workshop for the IFIP Working Group for Natural Language Processing and Knowledge Representation, the Computational Lexical Semantics of Verbs
University of Pennsylvania, USA

4-5 May 1995
IPSM ‘95, International Workshop on Industrial Parsing of Software Manuals
University of Limerick, Ireland
Tel: +353 61 202706, E-mail: sutcliffer@ul.ie

9-11 May 1996
Translation Studies: Unity in Diversity?
Dublin City University, Ireland
Jennifer Pearson, School of Applied Languages, Dublin City University, Dublin 9
E-Mail: pearsonj@dcu.ie

10-12 May 1995
2nd Symposium of Language Processing
Kasetsart University, Bangkok, Thailand
E-mail: pp@nontri.ku.ac.th

24-26 May 1995
International Conference on Cooperative Multimodal Communication, Theory and Applications
Eindhoven, The Netherlands
Tel: +31-13.66.23.80, E-Mail: denk@kub.nl

29-31 May 1995
NLULP5, 5th International Workshop on Natural Language Understanding and Logic Programming
Lisbon, Portugal
Tel: 351-1-295 3220, E-mail: gpl@fct.unl.pt

9 June 1995
Twente Workshop on Language Technology
Corpus-Based Approaches to Dialogue Modelling
Tel: +31 53 893680

E-mail: twlt@cs.utwente.nl

11-23 June 1995
The Fourth Annual CETH Summer Seminar, Center for Electronic Texts in the Humanities
Princeton University, USA
Tel: 908/932-1384, E-Mail: ceth@zodiac.rutgers.edu

20-22 June 1995
BISFAI '95, The Fourth Bar-Ilan Symposium on Foundations of Artificial Intelligence Focusing on Natural Languages and Artificial Intelligence
Ramat-Gan and Jerusalem, Israel
Tel: (972-3-) 715770/2, E-Mail: ariel@bimacs.cs.biu.ac.il

26-30 June 1995
ACL ‘95, 33rd Annual Conference of the Association of Computational Linguistics
MIT, Cambridge, Massachusetts, USA

28-29 June 1995
NLDB '95, First Workshop on Applications of Natural Language to Databases
Versailles, France
Prof. M. Bouzeghoub, Laboratoire PRiSM, Université de Versailles

30 June 1995
ACL/SIGDAT, Third Workshop on Very Large Corpora
MIT, Cambridge, Massachusetts, USA
E-mail: yarowsky@unagi.cis.upenn.edu

5-7 July 1995
Fourth International Conference on the Cognitive Science of Natural Language Processing
Dublin City University
E-Mail: alex@compapp.dcu.ie

5-7 July 1995
TMI95, 6th International Conference on Theoretical and Methodological Issues in Machine Translation
Leuven University, Belgium
E-mail: tmi@ccl.kuleuven.ac.be

10-21 July 1995
3rd European Summer School on Language and Speech Communication
Edinburgh, Scotland
Tel: 44 (0)31 650 4594, E-mail: elsnet@ed.ac.uk

11-13 July 1995
MT SUMMIT V
Luxembourg, European Commission
Tel: 4301-33423

11-15 July 1995

ACH-ALLC 95, Association for Computers and the Humanities and Association for Literary and Linguistic Computing
University of California, Santa Barbara, USA
Tel: 805/687-5003, E-mail: HCF1DAHL@ucsbuxa.ucsb.edu

2-4 August 1995
SNLP '95
The Second Symposium on Natural Language Processing
Central Plaza Hotel, Bangkok, Thailand
Tel: (662) 561-4621 Ext 182, E-mail: ak@nontri.ku.ac.th

14-25 August 1995
Formal Grammar
The Seventh European Summer School in Logic, Language and Information
Barcelona
E-mail: morrill@lsi.upc.es

19-21 August 1995
IJCAI-95 Workshop on Context in Natural Language Processing
Montreal, Canada
Tel: (313) 577-1667, E-Mail: lucja@cs.wayne.edu

9-13 September 1995
International Summer School on Contemporary Topics in Computational Linguistics
Tzigov Chark, Bulgaria
E-Mail: nicolas@aisb.edinburgh.ac.uk

14-16 September 1995
Recent Advances in Natural Language Processing
Velingrad, Bulgaria
Tel: +44 -131 650 2727 (Edinburgh), E-Mail: nicolas@aisb.edinburgh.ac.uk

20-23 September 1995
IWPT ’95, 4th International Workshop on Parsing Technologies
Prague/Karlovy Vary, Czech Republic
Tel: 31-13 66.30.60, E-mail: bunt@kub.nl

16-18 October 1995
Language Engineering Convention 1995
Queen Elizabeth II Conference Centre, London.

19-22 October 1995
International Symposium on Language, Logic and Computation
Tbilisi, Georgia
Tel: 44 131 650 4667, E-mail: tbilisi@cogsci.ed.ac.uk

Change of Address

If you change your address, please advise us on this form, or a copy, and send it to the following (this form can also be used to join the Group):

Mr J. D. Wigg
BCS-NLTSG
72 Brattle Wood
Sevenoaks
Kent TN13 1QU
UK Date:....../....../......

Name ....
Address ....
.... Postal Code...
Country .... E-Mail..
Tel.No .... Fax.No ....

We would like to know more about you and your interests and would be pleased if you would complete as much of the following questionnaire as you wish (please delete any unwanted words).

1. a. I am mainly interested in the computing/linguistic/user/all aspects of MT.
   b. What is/was your professional subject?..
   c. What is your native language?........ ....
   d. What other languages are you interested in? ....

2. What information in this Review (No.1, April '95) or any previous Newsletter, have you found:
   a. interesting? Year...... ....
   b. useful (i.e. some action was taken on it)? Year...... ....
   ....

3. Is there anything else you would like to hear about in the MT Review?
   ....
   ....
   ....

4. Would you be interested in participating in a project, such as:
   a. Reviewing MT books and/or MT/Multilingual software
   b. Researching/listing/reviewing public domain MT and MNLP software
   c. Designing/writing/reviewing MT/MNLP application software
   d. Designing/writing/reviewing general purpose (non application specific) MNLP procedures/functions for use in MT and MNLP programming
   e. Supervising MSc student MT/MNLP project
   Any other suggestions?..................... ....
   ....

Thank you for your time and assistance.