# Machine Translation — Ten Years On: An Overview of the Conference

## Derek Lewis

University of Exeter, UK

## Abstract

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 international conference on Machine Translation (MT) 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.

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 a PC. The evaluation and commercial performance of MT systems remains a key issue, alongside the wider question of who actually uses MT.

## 1 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 translation 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 word-processing systems.

## 2 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.

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 GPSG 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 shows 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 in 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 can 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.

# 3 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 word-processors. 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 (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 word-processors. Over 5000 users are claimed.

## 4 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.

## 5 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 have 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).

# 6 Proposals

Boh Wasyliw (De Montfort University) and Douglas Clarke (Cranfield University) presented a number of cases in which misunderstandings arising out of oral 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.