THE STATE OF MACHINE TRANSLATION IN EUROPE

John Hutchins
(University of East Anglia)

Abstract

This first half of this general survey covers MT and translation tools in use, including translators workstations, software localisation, and recent commercial and in-house MT systems. The second half covers the research scene, multilingual projects supported by the European Union, networking and evaluation.

In comparison with the United States and elsewhere, the distinctive features of activity in Europe in the field of machine translation and machine-aided translation are: (i) the development and popularity of translator workstations, (ii) the strong software localisation industry, (iii) the vigorous activity in the area of lexical resources and terminology, (iv) and the broad based research on language engineering supported primarily by European Union funds.

1. Translator's workstations

Recent years have seen great advances in the development and exploitation of support tools for translator. The four most widely used translator's workstations originate from Europe: Trados' Translation Workbench, IBM's TranslationManager, Eurolang's Optimizer, and STAR's Transit. In addition, Europe has been the centre for most of the background and current research on workstations: the TWB project funded by the European Union, and the Commission's own EURAMIS project.

For professional translators, the attraction of the workstation is the integration of tools from simple word processing aids (spelling and grammar checkers) to full automatic translation. The translator can choose to make use of whichever tool seems most appropriate for the task in hand. The vendors of these systems always stress that translators do not have to change their work patterns; the systems aim to increase productivity with translator-oriented tools which are easy to use and fully compatible with existing word processing systems. In facilities and functions, each offer similar ranges: multilingual split-screen word processing, terminology recognition, retrieval and management, translation memory (pre-translation based on existing texts), alignment software for users to create their own bilingual text databases, retention of original text formatting, and support for very wide range of European languages, both as source and target languages. Integration to MT systems is now provided by three of the workstations. In the case of Trados access is provided to the Transcend software from Intergraph; IBM Translation Manager links up with Systran; and Eurolang Optimzer with Logos.

After its disappointing experience with Eurotra, the European Commission has devoted most of its research support to the development of practical tools for translators and to the creation of essential lexical resources. Most of these projects will be described later. A major project was the TWB (Translator's Work Bench) project -- within the ESPRIT framework -- which began in 1989 and ended in 1994. The project, led by Triumph-Adler and involving 10 members from companies and universities, investigated the requirements of translators and proposed most of the features which are now commonplace in translator workstations: multilingual editor, document converters, access to lexica and terminology databases (e.g. Eurodicautom), access to MT systems (in this case, METAL), tools for term bank building, pre-translation and translation memory, and in particular a tool kit (System Quirk) for the analysis of texts and the development of lexical resource databases (thesauri, knowledge bases) from corpora and term banks. (A full description of the TWB project has recently been published: Kugler et al. 1995).

A second project in this area has been TRANSLEARN, an LRE project for an interactive corpus-based translation drafting tool (a prototype translation memory system) based on EU regulations and directives from the CELEX (European Union law) database. The languages involved have been English, French, Portuguese and Greek. A third project, this time funded by Volkswagen, is investigating the use of a domain knowledge base integrated with a linguistic database as a translation tool; the languages are German and Bulgarian.

The Translation Service of the Commission itself is now developing its own workstation: EURAMIS. The aim is to optimize the efficiency of the translation resources already available (e.g. the termbank Eurodicautom), to create a database of translated EU documents (as a 'translation memory'), and to provide easy access to MT systems. It will allow individual translators to develop their own tailor-made resources and facilities, with tools for text corpus management, glossary construction, and text alignment. A particular emphasis will be on the integration of MT and translation tools, including the mutual enrichment of Systran dictionaries and Eurodicautom lexical databases.

The European strength in the area of terminology continues. A recent list of termbanks from InfoTerm in Vienna reveals that over two thirds of the 100 terminological databases recorded are based in Europe. The links between terminologists and translators have been a marked feature on the European scene -- links which are now extending to the MT community. Terminology and the lexicon was the topic of an EAMT workshop in 1993 (Steffens 1995), and this August another EAMT workshop was held in conjunction with the international terminology conference in Vienna.

2. MT in use

As many observers have commented the take-up of MT systems in Europe has been much slower than expected; markets are small and fragmented, and professional translators remain hostile. The potential is enormous, but far from being exploited. MT systems are used primarily by large translation services and by multinational companies. Smaller organisations favour translation

workbenches, sometimes networked for sharing term databases and translation memories; and increasingly workstations are being considered by individual freelance translators. The cheaper PC-based systems are generally only of interest to those with occasional translation needs, and still not being purchased on the scale apparent in North America or Japan.

Some of the more notable recent installations in multinational companies to mention are: Ericsson, where the Logos system is providing 10% of translation needs (for producing manuals and documentation in French, German and Spanish); SAP, using Metal for German-English translation and Logos for English-French (totalling some 8 million words in the current year); and Siemens, providing a service based on Metal.

At the European Commission, the use of Systran continues to grow, now amounting to some 200,000 pages per year. The main users are non-linguist staff needing translations for information purposes (short-lived and/or repetitive documents), and drafts for writing documents in non-native languages. The increased use is attributable to improved interfaces and access tools. For translators the main development is EURAMIS.

However, perhaps the most distinctive feature of the European scene is the growth of companies providing software localisation. These services are acquiring considerable experience in the use of translation aids and MT systems (e.g. Logos, Metal and XL8). As a forum for the interchange of experience and the establishment of standards the Localisation Industry Standards Association was set up in 1990; the association publishes a newsletter (*LISA Forum*) and has produces a CD-Rom directory of products, standards and methods (*LISA Showcase*). A major centre for localisation is Ireland, which since 1994 has its own Software Localisation Group, holding conferences and workshops and recently setting up a Localisation Resources Center (with support from the Irish government and EU.)

3. Commercial systems

In comparison with the United States and Japan, there have been surprisingly few MT systems developed and manufactured by European organisations. Two come from the former Soviet Union. From St Petersburg comes the very successful PC-based Stylus Russian-English, English-Russian, and German-Russian systems oriented primarily to business correspondence. Sales have been good in both East and West Europe. From Kharkov (Ukraine) come the PARS systems for Russian and Ukrainian to and from English, with sales mainly in the Ukraine and East Europe.

In Germany, changes in organisation have affected the development of Metal in recent years, however it is now available as a client/server system on Unix workstations and PCs, and as a network service for occasional users. Best developed remains the German-English system, but other languages have now been released: German-Danish, French-English, English-Spanish, and work is reported on Catalan, Italian, Portuguese.

The most recent entrant from Germany is the Personal Translator, a joint product of IBM and von Rheinbaben & Busch, based on the LMT (Logic-Programming based Machine Translation) slot-grammar transfer-based system under development at IBM since 1985. LMT itself is available as an MT component for the IBM Translation Manager. The Personal Translator is a PC-based system and intended primarily for the non-professional translator (consultants, secretaries, technicians) and competing therefore with Globalink and similar products. At present the languages are German and English in both directions.

Other less sophisticated PC-based systems include the following: from Italy there is Hypertrans, developed initially for translating patents in Italian and English, then expanded to include other European languages (French and German, Spanish) and currently used to translate patent abstracts for the European Patent Office, and now sold for wider general purposes; from France there is the Al-Nakil system for translating between Arabic, French and English; and from Denmark there is the PC-based Winger system for Danish-English, French-English and English-Spanish, now also marketed in North America, and the TranSmart system for Finnish-English from Kielikone Ltd.

4. Custom-built systems

Both Winger and TranSmart were initially built for specific customers. In the case of TranSmart, this was developed originally as the Kielikone translation workstation for Nokia Telecommunications. Subsequently, versions were installed at other Finnish companies and the system is now being marketed more widely.

A similar story applies to GSI-Erli. This large language engineering company developed an integrated in-house translation system combining an MT engine and various translation aids and tools on a common platform AlethTrad. Recently it has been making the system available in customised versions for outside clients.

Custom-built MT has become a speciality of Cap Volmac Lingware Services, a Dutch subsidiary of the Cap Gemini Sogeti Group. Over the years this software company has constructed controlled-language systems for textile and insurance companies, mainly from Dutch to English.

The Dutch language has also been the focus of an in-house system built at Hook & Hatton, in this case initially for the translation of documents in chemical engineering, but now as a commercial MT service for industrial customers in other subject fields. Begun as a simple pattern matching method to find frequently recurring phrases, the system added a terminology database and simple grammar rules and is now an efficient low-cost low-quality MT system.

In recent years, probably the best known success story for custom-built MT is the PaTrans system developed for LingTech A/S to translate English patents into Danish. The system is based on methods and experience gained from the Eurotra project of the European Commission.

Nearly all the successful applications of MT and the in-house systems are based on the control of input texts. Research on this area of 'language engineering' has been strong in recent years in Europe. The SECC project in Leuven, supported by Siemens Nixdorf, Cap Gemini Innovation and Sietec is using MT methodology as the basis for a tool for writing in a controlled language, Simplified English; the Metal system is effectively translating English into the controlled language.

Earlier this year, the University of Leuven organised a conference on controlled languages, the first devoted exclusively to this theme (CLAW, 1996). Participation was international, but European research was strongly represented with 16 of the 22 contributions; the majority on controlled English, but also German and Swedish represented.

5. Research on MT

It is undeniable that MT research in Europe has quantitatively declined since the ending of the Eurotra project (officially in 1992, but effectively some years earlier). The failure to produce a working system has been attributed mainly to the over-emphasis on linguistic formalism, the neglect of the lexical databases, and the absence of industrial participation. Nevertheless, it is agreed that the project did create in Europe a strong research community in language engineering which was able to successfully collaborate cross-nationally.

Eurotra itself has however brought forth a number of continuations. One has been mentioned already: PaTrans. Another is the CAT2 system at Saarbrücken, an experimental platform (a stratificational transfer-based unification grammar system) which is the MT engine for the EU-funded (LRE) project ANTHEM. The prototype is to be a multilingual interface for natural language input and retrieval of medical diagnoses. The sublanguage system is based on the analysis of a Dutch and French corpus from the Belgian Army and on the widely used Systematised Nomenclature of Human and Veterinary Medicine.

Another by-product of Eurotra is the KIT-FAST system, an experiment in knowledge-based MT with an emphasis on tackling problems of anaphora and the representation of communication functions.

Two research groups are investigating MT systems where control of the input is the result of user-computer dialogue: monolinguals compose messages interactively in their own language and translation into an unknown target language is performed automatically. At the University of Manchester Institute of Science and Technology the experimental system is based on the 'pure' implementation of an example-based approach. At the University of Grenoble, the LIDIA system involves interactive disambiguation and reverse translation into the user's language.

Other research efforts are taking place at ISSCO (Geneva) within a unification grammar framework on a sublanguage system for avalanche warnings, at the Sharp Laboratories of Europe (Oxford) on the shake-and-bake model of MT, at the Institute for Information

Transmission Problems of the Russian Academy of Sciences (Moscow) on ETAP-3 -- a continuation of the Meaning-Text Model approach to MT begun in the 1970s, initially for French-Russian but now for English-Russian translation of electrical engineering and computer science texts.

Another project with East European antecedents is taking place in Berlin (Gesellschaft für Multilinguale Systeme) on a Russian-German system based on the Metal platform. The Russian analysis derives form previous research at the German Academy of Sciences and the German synthesis on systems developed for Metal. The aim is initially a SunSparc system.

The application of general-purpose natural language systems to automatic translation is becoming more and more common. The large LOLITA project at Durham University is aiming to develop a problem-and domain-independent natural language engineering core for multiple applications. One of these is translation, and reports have been given of an experiment involving Italian-English translation via a conceptual representation.

An outcome of the Core Language Engine project at Cambridge has been the SLT (Spoken Language Translator) developed as a prototype speech translation between Swedish and English in the domain of air travel planning, and first demonstrated in 1993. Subsequently, the team have shown the flexibility and portability of the architecture by the relatively rapid construction of a system for French and English speech translation in the same domain. The system operates with a statistically trainable processor producing quasi-logical representations.

The best known spoken language system under development in Europe is, of course, the Verbmobil project funded by the German Ministry of Research and Technology and involving research groups in a number of German universities. The aim is to produce a prototype dialogue system for negotiations for German and Japanese business people with English as a common dialogue language. The system is based on language-independent disambiguated representations incorporating speech act (dialogue) information and involving domain knowledge databases. A major effort has concentrated on the analysis of dialogue in the chosen domain.

6. Language Engineering projects

Since the ending of Eurotra, research funds from the European Union have been granted on a wide range of projects within the broad field of language engineering, which includes multilingual tools of all kinds as well as translation assistance in various contexts. Practical implementation and collaboration with industrial partners is emphasised throughout, as well as the need for general-purpose and re-usable products.

It is not possible to describe all those projects which involve multilinguality and translation. Only a few are highlighted. The ALEP project has been devoted to the development of a platform to support a wide range of research and technology activities related to natural language processing, including a general-purpose formalism and rule interpreter, and a text handling

system. (The most recent version is described in MTNI#14.) GRAAL was a parallel project for a re-usable grammar-writing formalism for text processing, including for computer-aided translation.

The development of lexical tools has been a major focus: GENELEX has defined a generic model for dictionary representations; DELIS is a tool to support lexicon building and management; MULTEXT is software for text corpora analysis and exploitation; and CRATER concentrates on bilingual corpus alignment.

More specifically translation oriented is the OTELO project with members including SAP (Germany), Lotus Development (Ireland), CST (Denmark) and Logos (Germany). The aim is to design a comprehensive automated translator's environment which combines at a single interface a variety of programs including MT, translation memory and other translation tools; and which will allow access from outside for new and potential users to try out MT over networks.

Some of the most recently approved projects are the following; which of them actually produce something worthwhile is yet to be seen. MABLe is to be a multilingual authoring system, guiding writers interactively to produce correspondence in a poorly known target language. MAY will provide multilingual access to yellow pages. MULTIMETEO is intended to generate weather forecasts in various languages from basic data supplied by meteorological computers. RECALL is to be a module for language learning which provides feedback translations for students. SPARKLE is a project to develop tools for syntactic analysis easily adapted to different languages and for semi-automatic lexical acquisition. SPEEDATA will provide continuous speech input for data entry in an Italian and German land registry. LINGUANET is to develop a multilingual communications system for police, based on experience with a controlled language Channel Tunnel (English and French) system. TREE will be provide multilingual access to a networked database of employment opportunities. TRADE is a project to translate social security reports in Italian, Spanish and English.

As before, there are a number of lexicon projects, nearly all multilingual (e.g. EUROWORDNET, INTERVAL), and projects to support the compilation and exploitation of corpora (e.g. PAROLE, SPEECHDAT). In the latter respect an important development is the establishment of the European Language Resources Association (ELRA) for the identification, collection, classification, validation, and exploitation of language resources (spoken and written), corpora, and linguistic models.

A number of European projects are tackling the automatic generation of natural language texts from databases. The APOLLO project is based on the CAT2 system mentioned earlier, with the aim of producing training documentation in French and English. DRAFTER is an experiment in the production of multilingual instruction documents for technical writers.

7. Networking MT

Europe was the location of one of the first networked MT service. This was the networking of Systran on the French Minitel system. Subsequently, other MT systems have become available over networks such as Internet: e.g. Logos, Metal and Globalink. Other efforts may be less familiar. TeleTaal is offering a PC-based word processing package with spelling and style checkers (in Dutch, French, German, Italian, and Spanish), which includes on-line support to electronic dictionaries and translation assistance through network links to the Globalink MT service. And there are experimental projects: MAITS is a EU-supported project to develop an interface to support access to MT services and translation memories, and TELELANG is investigating networked translation services (human translation as well as MT-based, terminology databases, linguistic resources, etc.

Finally, mention should be made of a service from British Telecom which produces summaries of texts in a wide range of languages and not restricted by subject. NetSumm is a statistics-based program which takes as input relatively short documents and produces gists in the form of extracted sentences. It is said to work best with newspaper articles and formal reports.

8. Evaluation

With the attention turning in Europe increasingly to usability of multilingual and translation systems there is inevitably much concern with valid and appropriate evaluation methods. An important vehicle is the EAGLES working group set up for establishing criteria for the evaluation and assessment of language engineering tools (which Margaret King reported on in AMTA-94). So far it has not looked at MT as such, having concentrated as yet on grammar checkers and writing aids.

Also important in this field is the TSNLP project (Essex University) which is seeking to establish test suites for natural language tools, including MT. And finally, we should remember that some of the most thorough examinations and evaluations of MT systems and translation aids have emanated from European organisations; most recently the two reports from Ovum Ltd. (London) on the global market and future prospects, and on the current translation technology available.

Sources

The sources of this survey are mainly conferences held in the last two years: particularly the MT Summit in Luxembourg in 1995, the Language Engineering Conventions in 1994 and 1995 (LEC 1994, 1995), the MT conference in Cranfield in 1995, and the Translating and the Computer conferences in 1995 (Aslib 1995), and the Sixth Theoretical and Methodological Issues in MT conference in Leuven, 1995 (TMI 1995). Other sources are various reports in *MT News International*, particularly those on MT in Europe by Colin Brace (MTNI#12), Paul Hearn (MTNI#14), Dorothy Senez (MTNI#14) and Jörg Schütz (MTNI#15).

Aslib (1995): Translating & the Computer 17. Papers from the Aslib conference held on 9th and 10th November 1995. London: Aslib, 1995.

CLAW (1996): Proceedings of the First International Workshop on Controlled Language Applications: CLAW96, 26-27 March 1996, Leuven, Belgium. [Leuven, 1996]

Cranfield (1994): International Conference: Machine translation ten years on, 12-14 November 1994, Cranfield University. [London: British Computer Society, 1994]

Kugler, M. et al. (1995): Translator's workbench: tools and terminology for translation and text processing. (Research Reports ESPRIT, project 2315: TWB.) Berlin: Springer.

LEC (1994): Language Engineering Convention, CNIT, La Défense, Paris, July 6-7, 1994. Abstracts. [Edinburgh, 1994]

LEC (1995): Second Language Engineering Convention, Queen Elizabeth II Conference Centre, London, 16-18 October 1995. Convention digest. London, 1995

MT Summit V (1995): MT Summit V. Proceedings, Luxembourg, July 10-13, 1995. [Luxembourg: SEMA Group, 1995]

Steffens, P. ed. (1995): Machine translation and the lexicon. Third International EAMT Workshop, Heidelberg, Germany, April 1993. Proceedings. Berlin: Springer, 1995.

TMI (1995): Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation: TMI95, July 5-7, 1995, Leuven, Belgium. [Leuven, 1995]