INTERNATIONAL JOURNAL OF TRANSLATION
Vol. 15, No. 1, Jan-June 2003

The Development and Use of Machine Translation Systems and Computer-based Translation Tools

JOHN HUTCHINS
(University of East Anglia, England)

INTRODUCTION

This survey of the present demand and use of computer-based translation software concentrates on systems designed for the production of translations of publishable quality, including developments in controlled language systems, translator workstations, and localisation; but it covers also the developments of software for non-translators, in particular for use with Web pages and other Internet applications, and it looks at future needs and systems under development. The final section compares the types of translations that can be met most appropriately by human and by machine (and computer-aided) translation respectively.

KEYWORDS: machine translation, computer-aided translation, translator workstations, multilingual systems

TYPES OF TRANSLATION DEMAND

When giving any general overview of the development and use of machine translation (MT) systems and translation tools, it is important to distinguish four basic types of translation demand. The first, and traditional one, is the demand for translations of a quality normally expected from human translators, i.e. translations of publishable quality - whether actually printed and sold, or whether distributed internally within a company or organisation. The second basic demand is for translations at a somewhat lower level of quality (and particularly in style), which are intended for users who want to find out the essential content of a particular document - and generally, as quickly as possible. The third type of demand is that for translation between participants in one-to-one communication (telephone or written correspondence) or of an unscripted presentation (e.g. diplomatic

NOTE : This article was written for presentation at a conference in Beijing (China) in June 1999. Some details about systems are therefore now out of date. For more recent information see the "Compendium of translation software" (www.eamt.org), or articles on my website: http://ourworld.compuserve.com/homepages/WJHutchins

exchanges.) The fourth area of application is for translation within multilingual systems of information retrieval, information extraction, database access, etc.

The first type of demand illustrates the use of MT for *dissemination.* It has been satisfied, to some extent, by machine translation systems ever since they were first developed in the 1960s. However, MT systems produce output which must invariably be revised or 'post-edited' by human translators if it is to reach the quality required. Sometimes such revision may be substantial, so that in effect the MT system is producing a 'draft' translation. As an alternative, the input text may be regularised (or 'controlled' in vocabulary and sentence structure) so that the MT system produces few errors which have to be corrected. Some MT systems have, however, been developed to deal with a very narrow range of text content and language style, and these may require little or no preparation or revision of texts.

In recent years, the use of MT systems for dissemination purposes has been augmented by developments in translation tools (e.g. terminology databases and translation memories), integrated in authoring and publishing processes. These 'translation workstations' are more attractive to human translators. Whereas, with MT systems translators see themselves as subordinate to the machine, in so far as they edit, correct or re-translate the output from a computer, with translation workstations (or workbenches) the translators are in control of computer-based facilities, which they can accept or reject as they wish.

The second type of demand - the use of MT for *assimilation* - has been met in the past as, in effect, a by-product of systems designed originally for the dissemination application. Since MT systems did not (and still cannot) produce high quality translations, some users have found that they can extract what they needed to know from the unedited output. They would rather have some translation, however poor, than no translation at all. With the coming of cheaper PC-based systems on the market, this type of use has grown rapidly and substantially.

With the third type - MT for *interchange* - the situation is changing quickly. The demand for translations of electronic texts on the Internet, such as Web pages, electronic mail and even electronic 'chat' lists, is developing rapidly. In this context, the possibility of human translation is out of the question. The need is for immediate translation in order to convey the basic content of messages, however poor the input. MT systems are finding a 'natural' role, since they can operate virtually or in fact in real-time and on-line and there has been little objection to the inevitable poor

quality. Another context for MT in personal interchange is the focus of much research. This is the development of systems for spoken language translation, e.g. in telephone conversations and in business negotiations. The problems of integrating speech recognition and automatic translation are obviously formidable, but progress is nevertheless being made. In the future - still distant, perhaps - we may expect on-line MT systems for the translation of speech in highly restricted domains.

The fourth type of MT application - as components of *information access* systems - is the integration of translation software into: (i) systems for the search and retrieval of full texts of documents from databases (generally electronic versions of journal articles in science, medicine and technology), or for the retrieval of bibliographic information; (ii) systems for extracting information (e.g. product details) from texts, in particular from newspaper reports; (iii) systems for summarising texts; and (iv) systems for interrogating non-textual databases. This field is the focus of a number of projects in Europe at the present time, which have the aim of widening access for all members of the European Union to sources of data and information whatever the source language.

HISTORICAL BACKGROUND

Systems for automatic translation have been under development for 50 years - in fact, ever since the electronic computer was invented in the 1940s there has been research on their application for translating languages (Hutchins 1986). For many years, the systems were based primarily on direct translations via bilingual dictionaries, with relatively little detailed analysis of syntactic structures. By the 1980s, however, advances in computational linguistics allowed much more sophisticated approaches, and a number of systems adopted an indirect approach to the task of translation. In these systems, texts of the source language are analysed into abstract representations of 'meaning', involving successive programs for identifying word structure (morphology) and sentence structure (syntax) and for resolving problems of ambiguity (semantics). Included in the latter are component programs to distinguish between homonyms (e.g. English words such as *light*, which can be a noun, and adjective or verb, and *solution*, which can be a mathematical or a chemical term) and to recognise the correct semantic relationships (e.g. in *The driver of the bus with a ye/low coat).* The abstract representations are intended to be unambiguous and to provide the basis for the generation of texts into one or more target languages. There have in fact been two basic 'indirect' approaches. In one the abstract representation is designed to be a kind of language-independent

'interlingua', which can potentially serve as an intermediary between a large number of natural languages. Translation is therefore in two basic stages: from the source language into the interlingua, and from the interlingua into the target language. In the other indirect approach (in fact, more common approach) the representation is converted first into an equivalent representation for the target language. Thus there are three basic stages: analysis of the input text into an abstract source representation, transfer to an abstract target representation, and generation into the output language.

Until the late 1980s, systems of all these kinds were developed, and it is true to say that all current commercially available systems are also classifiable into these three basic system types: direct, interlingual and 'transfer'. The best known of the MT systems for mainframe computers are in fact essentially of the 'direct translation' type, e.g. the Systran, Logos and Fujitsu (Atlas) systems. They are however improved versions of the type; unlike their predecessors, they are highly modular in construction and easily modifiable and extendable. In particular, the Systran system, originally designed for translation only from Russian into English, is now available for a very large number of language pairs: English from and into most European languages (French, German, Italian, Spanish, Portuguese), Japanese, Korean, etc. Logos, originally marketed for German to English, is also now available for other languages: English into French, German, Italian and Spanish, and German into French and Italian. The Fujitsu ATLAS system, on the other hand, is still confined to translation between English and Japanese (in both directions).

Among the most important of the mainframe 'transfer' systems was METAL, supported for most of the 1980s by Siemens in Germany. However, it was only at the end of the decade that METAL came onto the market, and sales were poor. During the 1990s, rights to METAL have been transferred to two organisations (GMS and LANT) in a complex arrangement. But the best known systems adopting the 'transfer' approach were research projects: Ariane at GETA in Grenoble, an MT project going back to the 1960s, and Eurotra funded by the Commission of the European Communities. There were hopes that Ariane would become the French national system, and there were plans to incorporate it in a translator's workstation for Eurolang (see below), but in the end nothing came of them. As for Eurotra, it was undoubtedly one of the most sophisticated systems, but after involving some hundred of researchers in most countries of Western Europe for almost a decade, it failed to produce the working system that the sponsors wanted. It had been hoped that Eurotra would eventually

replace the Systran systems that the Commission had acquired and was developing internally. In the late 1980s, Japanese governmental agencies began to sponsor an interlingua system for Asian languages, involving co-operation with researchers in China, Thailand, Malaysia and Indonesia. However, this project too has so far not produced a system after a decade of work. (For surveys of MT research and development in 1980s and early 1990s see Hutchins 1993, 1994.)

GOVERNMENTAL AND NON-COMMERCIAL USE

The earliest installations of MT systems were in national and international governmental and military translation services - primarily because they could afford the costs of the computer hardware required. The US Air Force introduced Systran in 1970 for translating Russian military scientific and technical documentation into English. Although some documents were edited, much of the output was passed to recipients without revision; over 90% accuracy for technical reports is claimed. The National Air Intelligence Center, which took over the service from the USAF, now produces translations (many unedited) for a wide range of US government organisations (Pedtke 1997). As well as Russian-English it has available systems from Systran for translating Japanese, Chinese and Korean into English, and under development with Systran is a system for SerboCroat into English.

In Europe, the largest translation service is that of the European Commission, and was one of the first organisations to install MT. It began in 1976 with the Systran system for translating from English into French. In subsequent years, versions were developed for many other language pairs, covering the needs for translation among the European Union languages. While the translation of many legal texts continues to be done by human translators, the Systran systems are used increasingly not only for the translation of internal documents (with or without post-editing) but also as rough versions for the assistance of administrators when composing texts in non-native languages (Senez 1996).

PRODUCTION OF TECHNICAL DOCUMENTATION

Until the 1990s the normal assumption was that MT systems were intended to be used for the production of documentation of publishable quality, primarily but not exclusively of a scientific and technical nature. The assumption was, in other words, that MT systems were to be used in conditions where otherwise human translators would be employed with expertise in the subjects concerned. Evidently, the actual quality of MT

output was inadequate for direct use. It had to be extensively revised before it could be published, and translators were therefore employed as 'post-editors'. In these circumstances, the use of MT became a matter of economics. It was viable only if overall quality and speed could be achieved at lower cost than the employment of human translators.

Although today there are other uses for MT, as we have already indicated, this application remains the most important, particularly for the vendors and developers of the larger 'mainframe'-type systems (Systran and Logos). The main customers and users are the multinational companies exporting equipment in the global market (Vasconcellos 1993; Brace et al. 1995). The need here is for translation of promotional and technical documentation. In the latter case, technical documents are often required in very large volumes: a set of operational manuals for a single piece of equipment may amount to several thousands of pages Furthermore, there can be frequent revisions with the appearance of new models. In addition, there must be consistency in translation: the same component must be referred to and translated the same way each time. This scale of technical translation is well beyond human capacity. Nevertheless, in order to be most cost-efficient, a MT system should be well integrated within the overall technical documentation processes of the company: from initial writing to final publishing and distribution. Systems developed for the support of technical writers - not just assistance with terminology, but also on-line style manuals and grammar aids - are now being linked seamlessly into translation and publishing processes.

There are numerous examples of the successful and long-term use of MT systems by multinationals for technical documentation. One of the best known is the application of the Legos systems at the Lexi-Tech company in New Brunswick, Canada; initially for the translation into French of manuals for the maintenance of naval frigates, the company has built up a service undertaking many other large translation projects. Also using Logos are Ericsson, Osram, Oce Technologies, SAP and Corel. Systran has many large clients: Ford, General Motors, Aerospatiale, Berlitz, Xerox, etc. The METAL German-English system has been successfully used at a number of European companies: Boehringer Ingelheim, SAP, Philips, and the Union Bank of Switzerland.

A pre-requisite for successful MT installation in large companies is that the user expects a large volume of translation within a definable domain (subjects, products, etc.) The financial commitment to a terminology database and to dictionary maintenance must be justifiable. Whether produced automatically or not, it is desirable for company documentation

to be consistent in the use of terminology. Many companies in fact insist upon their own use of terms, and will not accept the usage of others. To maintain such consistency is almost impossible outside an automated system. However, it does mean that before an MT system can be installed, the user must have already available a well-founded terminological database, with authorised translation equivalents in the languages involved, or - at least - must make a commitment to develop the required term bank.

For similar reasons, it is often desirable if the MT system is to produce output in more that one target language. Most large-scale MT systems have to be customised, to a greater or lesser extent, for the kind of language found in the types of documents produced in a specific company. This can be the addition of specific grammatical rules to deal with frequent sentence and clause constructions, as well as the inclusion of specific rules for dealing with lexical items, and not just those terms unique to the company. The amount of work involved in such customisation may not be justifiable unless output is in a number of different languages.

CONTROLLED LANGUAGE AND DOMAIN-SPECIFIC SYSTEMS

In these circumstances, however, it has often been found feasible to introduce a greater degree of control. One of the earliest and best known examples is the application of the Systran system by the Xerox Corporation. At Xerox technical authors are obliged to compose documents in what is called Multinational Customized English, where not only the use of specific terms is laid down but also the construction of sentences (Elliston 1979). The advantages of this approach are: the avoidance of ambiguities in the input which the MT system cannot deal with adequately, the consequential better quality output, the faster production of technical documents simultaneously in a number of different languages, and (not least) the production of more easily comprehensible English documents. These advantages have been recognised by other multinational companies, and the use of 'controlled languages' is increasing: for example, the Caterpillar Corporation has devised its own form of English to facilitate translation in a knowledge-based MT system being developed for it at the Carnegie-Mellon University (Mitamura and Nyberg 1995). There are some companies offering to build 'controlled' language MT systems for specific clients. The oldest established - and the pioneer in this approach - is the Smart Corporation, New York. Systems have been developed by Smart for a number of major clients: Citicorp, Chase, Ford, General Electric, etc. Each incorporates a system for 'normalising' English documents. This

system component is considered so crucial to success that the actual translation process is regarded as virtually a 'by-product' (Lee 1994). There are Smart systems translating into French, German, Greek, Italian, Japanese, and Spanish. The largest Smart installation, perhaps, is the system designed for the Canadian Ministry of Employment, where it has been used for many years to translate information about job advertisements and similar documentation.

In Europe, the Cap Volmac company in the Netherlands and the LANT company in Belgium offer similar services, building for various clients specialised translation systems utilising their own software for controlled languages. Cap Volmac Lingware Services is 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 (Van der Steen and Dijenborgh 1992). However, possibly 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 (0rsnes et al. 1996)

These last examples of systems illustrate that a growing number of companies and organisations are developing their own MT facilities, as opposed to purchasing commercial systems. This has been a feature from early days. The successful Météo system in Canada for translating weather forecasts from English into French (and later from French into English) was effectively a customer-specific system - in this case the Canadian Environment service. It may be noted that a variant of the Météo software was successfully operated during the Olympic games in Atlanta (Chandioux and Grimaila 1996). Météo is an example of a 'sublanguage' system, i.e. designed for to deal with the particular language of meteorology.

Another example of a custom-built system is TITUS, a highly constrained 'sublanguage' system for translating abstracts of documents of the textile industry from and into English, French, German, and Spanish, in regular use since 1970. Better known are the two customer-specific systems for translating between English and Spanish built at the Pan American Health Organization in Washington - designed and developed by workers in the organisation itself. These highly successful systems (now also available to users outside PAHO) are general-purpose systems, not constrained in vocabulary or text type, although obviously the dictionaries are strongest in the health-related social science fields (Leon and Aymerich 1997).

In the 1990s there have been a number of other examples. In Finland, the Kielikone system was developed originally as a workstation for Nokia Telecommunications. Subsequently, versions were installed at other Finnish companies and the system is now being marketed more widely (Arnola 1996). A similar story applies to GSI-Erli. This large language engineering company developed an integrated in-house translation system combining a 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 (Humphreys 1996).

On a smaller scale, but equally successful, has been the system developed by the translation service of a small British company Hook and Hatton. In this case, the need was for translation of chemical texts from Dutch into English (Lewis 1997). The designer began by simple pattern matching of phrases, and gradually built in more syntactic analysis as and when results were justifiable and cost-effective.

Based on experience over many years in developing knowledge-based MT and experimenting with speech translation and corpus-based methods, members of the group at Carnegie-Mellon University have developed an architecture for the rapid production of usable MT systems for specific clients in some less common languages, such as SerboCroat and Haitian Creole (Frederking et al. 1997). There is no pretence of high quality, merely 'usefulness' for languages otherwise inaccessible.

Another example of custom-built MT in a specialised area is the program developed for TCC Communications at the Simon Fraser University for translating closed captions on television programs (Toole et al. 1998). Not only are there time constraints - translation must be in real-time - but also there are the challenges of colloquialisms, dialogue, robustness, and paucity of context indicators. The system, at present running live for English into Spanish, demanded techniques otherwise found mainly in Internet applications (see below.)

In Japan, there are further examples of custom-built systems. The Japan Information Centre of Science and Technology translates abstracts of Japanese scientific and technical articles into English. In the late 1980s it assumed responsibility of the Mu Japanese-English MT system developed at the University of Kyoto. From this, it now has one of the largest MT operations in Japan (O'Neill-Brown 1996). Other custom-built systems of significance in Japan are the SHALT system developed by IBM Japan for its own translation needs, the ARGO system developed by CSK in Tokyo for translating Japanese stock market reports into English, and the NHK system for translating English news articles into Japanese.

TRANSLATION WORKSTATIONS

In the 1990s, the possibilities for large-scale translation broadened with the appearance on the market of translation workstations (or translator workbenches). The original ideas for integrating various computer-based facilities for translators at one place go back to the early 1980s, in particular with the systems from ALPS. Translation workstations combine multilingual word processing, means of receiving and sending electronic documents, OCR facilities, terminology management software, facilities for concordancing, and in particular 'translation memories'. The latter is a facility that enables translators to store original texts and their translated versions side by side, i.e. so that corresponding sentences of the source and target are aligned. The translator can thus search for a phrase or even full sentence in one language in the translation memory and have displayed corresponding phrases in the other language. These may be either exact matches or approximations ranked according to closeness.

It is often the case in large companies that technical documents, manuals, etc. undergo numerous revisions. Large parts may remain unchanged from one version to the next. With a translation memory, the translator can locate and re-use already translated sections. Even if there is not an exact match, the versions displayed may be usable with minor changes. There will also be access to terminology databases, in particular company-specific terminology, for words or phrases not found in the translation memory. In addition, many translator workstations are now offering full automatic translations using MT systems such as Systran, Logos, and Transcend. The translator can choose to use them either for the whole text or for selected sentences, and can accept or reject the results as appropriate (Heyn 1997).

There are now four main vendors of workstations: Trados (probably the most successful), STAR AG in Germany (Transit), IBM (the TranslationManager), and LANT in Belgium (the Eurolang Optimizer, previously sold by SITE in France). The translation workstation has revolutionised the use of computers by translators. They have now a tool where they are in full control. They can use any of the facilities or none of them as they choose. As always, the value of each resource depends on the quality of the data. As in MT systems, the dictionaries and terminology databases demand effort, time and resources. Translation memories rely on the availability of suitable large corpora of authoritative translations - there is no point in using translations which are unacceptable (for whatever reason) by the company or the client.

Although widely used by administrators within the European Commission, the full-scale MT system Systran is relatively little used by the Commission's professional translators. For them, the translation service is developing its own workstation facility, EURAMIS, i.e. European Advanced Multilingual Information System (Theologitis 1997). This combines access to the Commission's own very large multilingual database (Eurodicautom), the dictionary resources of Systran, facilities for individual and group terminology database creation and maintenance (using Trados' MultiTerm software), translation memory (again for individuals and groups), access to CELEX (the full-text database of European Union legislation and directives), software for document comparison (to detect where changes have taken place), and also, of course, access to the Systran MT systems themselves. The latter are now available from English into Dutch, French, German, Greek, Italian, Portuguese, and Spanish; from French into Dutch, English, German, Italian, and Spanish; from Spanish into English and French; and from German into English and French. The whole EURAMIS system is linked to other facilities such as authoring tools (spelling, grammar and style checkers, and multilingual drafting aids), the internal European Commission administrative network, and to outside resources on the Internet.

LOCALISATION OF SOFTWARE

One of the fastest growing areas for the use of computers in translation is in the industry of software localisation. Here the demand is for supporting documentation to be available in many languages at the time of the launch of new software. Translation has to be done quickly, but there is much repetition of information from one version to another. MT and, more recently, translation memories in translation workstations are the obvious solution (Schaeler 1996). Among the first in this field was the large software company SAP AG in Germany. They use two MT systems: METAL for German to English translation, and Logos for English to French, and plan to introduce further systems for other language pairs.

Most localisation, however, is based on the translation memory and workstation approach. Typical examples are Corel, Lotus, and Canon. It is interesting to note that much of this localisation activity is based in Ireland - thanks to earlier government and European Union support for the computer industry. However, localisation is a multi-national and global industry, with its own organisation (Localization Industry Standards Association, based in Geneva) holding regular seminars and conferences in all continents (For details see *LISA Forum Newsletter)*

Localisation companies have been at the forefront of efforts in Europe to define standardised lexical resource and text handling formats, and to develop common network infrastructures. This is the OTELO project coordinated by Lotus in Ireland, with other members such as SAP, Logos, and GMS. The need for a general translation environment for a wide variety of translation memory, machine translation and other productivity tools is seen as fundamental to the future success of companies in the localisation industry.

SYSTEMS FOR PERSONAL COMPUTERS

Software for personal computers began to appear in the early 1980s (with the Weidner MicroCAT system becoming particularly successful). Nearly all the main Japanese computer companies produced systems for translation to and from English, e.g. the PIVOT system from NEC, the ASTRANSAC system from Toshiba, HICATS from Hitachi, PENSEE from Oki and DUET from Sharp.

Outside Japan, systems for personal computers began to appear a little earlier, but from relatively few companies. The first American systems came in the early 1980s from ALPS and from Weidner. The ALPS products were intended primarily as aids for translation, providing tools for accessing and creating terminology resources but they did include an interactive translation module. Although at first sold with some success, the producers concluded by the end of the decade that the market was not yet ready and the products were in effect withdrawn. Instead, ALPS turned itself into a translation service (ALPNET), using its own tools internally. By contrast, Weidner sold a full translation system in a growing number of language pairs (English, French, German, Spanish), and the business flourished. Weidner produced two versions of its systems: MicroCat for small personal computers, and MacroCat for larger minicomputers or workstations. The company was then purchased by a Japanese company Bravis, a Japanese version was marketed, but soon afterwards the owner decided that the MT market for personal computers was still undeveloped and the business was sold. MicroCat disappeared completely, but MacroCat was purchased by Intergraph, who modified and developed it for its range of publishing software and sold it later as Transcend - recently Transcend was acquired by Transparent Language Inc. (For these developments see Hutchins 1993, 1994).

At the end of the 1980s, most of the commercial systems on the market today appeared. First came the PC-Translator systems (from Linguistic Products, based in Texas) for low-end personal computers. Over the years,

many language pairs have been produced and marketed, apparently successfully as far as sales are concerned. Next came Globalink with systems for French, German and Spanish to and from English. (There was also a Russian-English system deriving essentially from the original owner's experience on the 1960s Georgetown project.) Within a few years, Globalink merged with MicroTac, a company which had been very successful in selling its cheap Language Assistant series Of PC software - essentially automatic dictionaries, with minimal phrase translation facility. In the early 1990s, Globalink produced its now well-known 'Power Translator' series for translation of English to and from French, German and Spanish, and recently Globalink has marketed the more advanced 'Telegraph' series of translation software products, and Globalink itself was acquired by Lernout & Hauspie, a leading speech technology company.

Since the beginning of the 1990s, many other systems for personal computers have appeared. For Japanese and English there are now also LogoVista from the Language Engineering Corporation, and Tsunami and Typhoon from Neocor Technologies (also now owned by Lernout & Hauspie). From the former Soviet Union - where particularly in the 1960s and 1970s there was very active research on MT - we have now Stylus (recently renamed ProMT) and PARS, both marketing systems for Russian and English translation; Stylus also for French, and PARS also for Ukrainian. Other PC-based systems from Europe include: Hypertrans for translating between Italian and English; the Winger system for Danish-English, French-English and English-Spanish, now also marketed in North America; and TranSmart, the commercial version of the Kielikone system, for Finnish-English translation.

Vendors of older mainframe systems (Systran, Fujitsu, Metal, Logos) are being obliged to compete by downsizing their systems; many have done so with success, managing to retain most features of their mainframe products in the PC-based versions. Systran Pro and Systran Classic, for example, are Windows-based versions of the successful system developed since the 1960s for clients worldwide in a large range of languages; the large dictionary databases offered by Systran give these systems clear advantages over most other PC products. Both Systran Classic (for home use) and Systran Pro (for use by translators) are now sold for under a five hundred dollars in many language pairs: English-French, English-German, English-Spanish; and for English to Italian and Japanese to English. The publishing company Langenscheidt acquired rights to sell a version of METAL, in collaboration with GMS (Gesellschaft für Multilinguale Systeme, now owned by Lernout & Hauspie) - the system is called

'Langenscheidt T1' and offers various versions for German and English translation. Also from Germany is the Personal Translator, a joint product of IBM and von Rheinbaben & Busch, based on the LMT (i.e. Logic-Programming based Machine Translation) transfer-based system under development since 1985. LMT itself is available as a MT component for the IBM TranslationManager. Both Langenscheidt T1 and the Personal Translator are intended primarily for the non-professional translator, competing therefore with Globalink and similar products. (For these developments see proceedings of MT conferences: AMTA, EAMT, MT Summit, and *MT News International.)*

Sales of commercial PC translation software have shown a dramatic rise. There are now estimated to be some 1000 different MT packages on sale (when each language pair is counted separately.) The products of one vendor (Globalink) are present in at least 6000 stores in North America alone; and in Japan one system (Korya Eiwa from Catena, for English-Japanese translation) is said to have sold over 100,000 copies in its first year on the market. Though it is difficult to establish how much of the software purchased is regularly used (some cynics claim that only a very small proportion is tried out more than once), there is no doubting the growing volume of 'occasional' translation, i.e. by people from all backgrounds wanting renderings of foreign texts in their own language, or wanting to communicate in writing with others in other languages, and who are not deterred by poor quality. It is this latent market for low-quality translation, untapped until very recently, which is now being discovered and which is contributing to massive increases in sales of translation software.

MT ON THE INTERNET

At the same time, many MT vendors have been providing network-based translation services for on-demand translation, with human revision as optional extras. In some cases these are client-server arrangements for regular users; in other cases, the service is provided on a trial basis, enabling companies to discover whether MT is worthwhile for their particular circumstances and in what form. Such services are provided, for example, by Systran, Logos, Globalink, Fujitsu, JICST and NEC.

Some companies have now been set up primarily for this purpose: LANT in Belgium is a major example, based on its rights to develop the METAL system and on the Eurolang Optimizer, which it also markets (Caeyers 1997). Its speciality is the customisation of controlled languages for use with its MT and translation memory systems. In late 1997 it launched

its multilingual service for the translation of electronic mail, Web pages and attached files. And in Singapore, there is MTSU (Machine Translation Service Unit of the Institute of Systems Science, National University of Singapore), using its own locally-developed systems for translation from English into Chinese, Malay, Japanese and Korean (with Chinese its main strength) and with editing by professional translators. The service is providing large scale translation over the Internet for many customers world wide (mainly multinational organisations), and including much of the localisation needs for software companies in the Chinese-language markets *(LISA Forum Newsletter* 4(3), August 1995, p. 12.)

A further sign of the influence of Internet is the growing number of MT software products for translating Web pages. Japanese companies have led the way: nearly all the companies mentioned above have a product on this lucrative market; they have been followed quickly elsewhere (e.g. by Systran, Globalink, Transparent Language, LogoVista). As well as PC software for translating Web pages, we are now seeing Internet services adding translation facilities: the most recent example is the availability on AltaVista of versions of Systran for translating French, German and Spanish into and from English - with what success or user satisfaction it is too early to say (Yang and Lange 1998).

Equally significant has been the use of MT for electronic mail and for 'chat rooms'. Two years ago CompuServe introduced a trial service based on the Transcend system for users of the MacCIM Support Forum. Six months later, the World Community Forum began to use MT for translating conversational e-mail. Usage has rocketed (Flanagan 1996). Most recently, CompuServe introduced its own translation service for longer documents either as unedited 'raw' MT or with optional human editing. Soon CompuServe will offer MT as a standard for all its e-mail. As for Internet chat, Globalink has joined with Uni-Verse to provide a multilingual service.

The use is not simple curiosity, although that is how it often begins. CompuServe records a high percentage of repeat large-volume users for its service, about 85% for unedited MT - a much higher percentage than might have been expected. It seems that most is used for assimilation of information, where poorer quality is acceptable. The crucial point is that customers are prepared to pay for the product - and CompuServe is inundated with complaints if the MT service goes down!

It is clear that the potential for MT on, via and for the Internet is now being fully appreciated - no company can afford to be left behind, and all the major players have ambitious plans, e.g. Lernout & Hauspie (McLaughlin and Schwall 1998), which has now acquired MT systems

from Globalink, Neocor and AppTek as well as the old METAL system (from GMS).

FUTURE NEEDS AND DEVELOPMENTS

Despite the recent growth of systems for personal computers and of Internet services, it is still true to say that there is nothing yet really suitable for the independent professional translator, i.e. for those not working for large companies or in translation organizations. It is known that some translators have tried to apply commercial PC-based software to their needs, but the amount of adaptation required and the generally poor output has made them unsatisfactory and uneconomic. More suitable for the independent translator would be a cost-effective translation workstation. However, current workstations on the market are still too expensive for the individual translator. Although there is promise of low-cost computer tools for this potentially large market - e.g. terminology and concordancing software, and perhaps alignment software - there is no doubt that this segment is not being covered as well as many other areas.

Another area at present poorly served is the need for reliable but low-cost translation of documents into unknown foreign languages where users do not want to engage expert bilingual translators. There is no problem with translation into recipients' own languages - PC systems can give adequate 'rough' versions for users to get some idea of the basic message - but for translation into an unknown language there are still no solutions. There have been recently some cheap Japanese products which serve this specific 'foreign language authoring' demand in the case of writing business letters (based on standard phrases and document templates), but for other areas and for longer documents, where there is less 'stereotyping', there is nothing as yet. For translation into another language unknown (or poorly known) by the sender, what is really required is software which can be relied upon to provide good quality output (and most PC products are not good enough). A number of research groups are investigating interactive systems, where the sender composes an MT-friendly version of a letter or document in collaboration with the computer. With a sufficiently 'normalised' input text, the MT system can guarantee grammatically and stylistically correct output. As yet, however, this work (e.g. at GETA in France) is still at the laboratory stage (Boitet and Blanchon 1995).

The same is true for software combining MT with information access, information extraction, and summarisation software. There are no commercial systems yet on the market; developments are still at the research stages. The potential and the demand have been recognised: for example, in

recent years, most research funds of the European Union have been focused not on MT or 'pure' natural language processing (as it was during the 1980s), but on projects for multilingual tools with direct applications in mind; many involve translation of some kind, usually within a restricted subject field and often in controlled conditions (Hutchins 1996; Schütz 1996). As just one example, the AVENTINUS project is developing a system for police forces in the area of drug control and law enforcement: information about drugs, criminals and suspects will be available on databases accessible in any of the European Union languages.

There is growing interest in such multilingual applications worldwide. The application that has received most attention has been 'cross-language information retrieval', i.e. software enabling users to search foreign language databases in their own languages. As yet most work has focussed on the construction and operation of appropriate translation dictionaries, for the matching of query words against words or phrases in document databases (Bian and Chen 1998, Oard 1998) - although the provision of software for fast translation of original texts into the enquirer's own language is naturally also envisaged (McCarley and Roukos 1998). Clearly it will not be long before commercial software is available for this application.

The future application that is probably most desired by the general public is the translation of spoken language. But, from a commercial (and even research) perspective, the prospects for automatic speech translation are still distant (Krauwer et al. 1997). It was only in the 1980s that developments in speech recognition and synthesis made spoken language translation a feasible objective. In Japan a joint government and industry company ATR was established in 1986 near Osaka, and it is now one of the main centres for automatic speech translation. The aim is to develop a speaker-independent real-time telephone translation system for Japanese to English and vice versa, initially for hotel reservation and conference registration transactions. Other speech translation projects have been set • up subsequently. The JANUS system is a research project at Carnegie-Mellon University and at Karlsruhe in Germany. The researchers are collaborating with ATR in a consortium (C-STAR), each developing speech recognition and synthesis modules for their own languages (English, German, Japanese). (One by-product of this research was mentioned earlier: the rapid-deployment project for custom-built systems in less-common languages.) The fourth major effort in speech translation is the long-term VERBMOBIL project funded by the German Ministry for Research and Technology which began in May 1993. The aim is a portable aid for

business negotiations as a supplement to users' own knowledge of the languages (German, Japanese, English). Numerous German university groups are involved in fundamental research on dialogue linguistics, speech recognition and MT design; a prototype is nearing completion, and a demonstration product is targeted for early in the next century.

Speech translation is probably at present the most innovative area of computer-based translation research, and it is attracting most funding and the most publicity. However, few experienced observers expect dramatic developments in this area in the near future - the development of MT for written language has taken many years to reach the present stage of widespread practical use in multinational companies, a wide range of PC based products of variable quality and application, growing use on networks and for electronic mail. Despite today's high profile for written-language MT, researchers know that there is still much to be done to improve quality. Spoken-language MT has not yet reached even the stage of real-time testing in non-laboratory settings.

COMPARISON OF HUMAN AND MACHINE TRANSLATION

From this survey it should be apparent that the application of computers to the task of translating natural languages has not been and is unlikely to be a threat to the livelihood of professional translators. Those skills which the human translator can contribute will continue always to be in demand. There is no prospect, for example, that machine translation could ever attempt the translation of literary or legal texts. By contrast, for the rough translation of electronic texts on the Internet there is no rivalry for machine translation - human translators cannot compete in terms of speed, even if they were prepared to undertake poor quality translation of ephemeral material.

We may compare the relative merits of human and machine translation according to the categories of need and use outlined at the beginning of this paper. As far as the dissemination function (production of publishable translations) is concerned, human translation is more satisfactory and less costly overall whenever it is a question of translating one particular text in a unique subject domain (whether scientific, technical, medical, legal or literary). Machine translation demands the costly investment of dictionary maintenance and updating and the costly involvement of post-editing. This can be justifiable (i.e. cost-effective) only when large volumes of documentation within a particular domain are being translated. It is even more justifiable if translation is into more than one target language (when pre-editing and/or vocabulary and grammar control of original texts is

possible), and when there is considerable repetition. For such tasks, the human translator would be overwhelmed by the scale of the task, by the boring repetitiveness and by the need to maintain terminological consistency. By contrast, the computer can handle large volumes and can automatically maintain consistency. In brief, machine translation is ideal for large scale and/or rapid translation of (boring) technical documentation, (highly repetitive) software localisation manuals, and real-time translation of weather reports. The human translator is (and will remain) unrivalled for non-repetitive linguistically sophisticated texts (e.g. in literature and law).

For the translation of texts for assimilation, where the quality of output can be poorer than that for texts to be published, it is clear that machine translation is an ideal solution. Human translators are not prepared (and resent being asked) to produce 'rough' translations of scientific and technical documents that may be read by only one person who wants to merely find out the general content and information and is unconcerned whether everything is intelligible or not, and who is certainly not deterred by stylistic awkwardness or grammatical errors. Of course, they might prefer to have output better than that presently provided by most MT systems, but if the only alternative option is no translation at all then machine translation is fully acceptable.

For the interchange of information, there may still in the future continue to be a role for the human translator in the translation of business correspondence (particularly if the content is sensitive or legally binding). But for the translation of personal letters, MT systems are likely to be increasingly used; and, for electronic mail and for the extraction of information from Web pages and computer-based information services, MT is the only feasible solution.

For spoken translation, by contrast, there will be a continuing market for the human translator. There is surely no prospect of automatic translation replacing the interpreter of diplomatic and business exchanges. Although there has been research on the computer translation of telephone enquiries within highly constrained domains, and future implementation can be envisaged in this area, for the bulk of telephone communication there is unlikely to ever be any substitute for human translation.

Finally, MT systems are opening up new areas where human translation has never featured: the production of draft versions for authors writing in a foreign language, who need assistance in producing an original text; the on-line translation of television subtitles, the translation of information from databases; and no doubt, more such new applications will appear in

the future. In these areas, as in others mentioned, there is no threat to the human translator because they were never included in the sphere of professional translation. There is no doubt that MT and human translation can and will co-exist in harmony and without conflict.

REFERENCES

AMTA 1996. Expanding MT horizons: Proceedings of the Second Conference of the Association for Machine Translation in the Americas, 2-5 October 1996, Montreal, Canada. [Washington, D.C.: AMTA]

AMTA 1998. Machine translation and the information soup: Third Conference of the Association for Machine Translation in the Americas... Langhorne. PA, USA, October 1998: Proceedings, ed. D.Farwell, L.Gerber [and] E.Hovy. Berlin: Springer-Verlag, 1998. (Lecture Notes in Artificial Intelligence 1529)

Arnola, H. 1996. Kielikone machine translation technology and its perspective on the economics of machine translation. In: EAMT Workshop (1996), 73-88.

Bian, G.W. and Chen, H. H. 1998. Integrating query translation and document translation in a cross-language information retrieval system. In: *AMTA (1*998), 250-265.

Boitet, C. and Blanchon, H. 1995. Multilingual dialogue-based MT for monolingual authors: the LIDIA project and a first mockup. *Machine Translation* 9(2), 99-132.

Brace, C., Vasconcellos, M. and Miller, L.C. 1995. MT users and usage: Europe and the Americas. *MT News International* 12 (October 1995), 14-19.

Caeyers, H. 1997. Presentation of LANT technology. In: *MT Summit (1997*\ 253-254.

Chandioux, J. and Grimaila, A. 1996. 'Specialized' machine translation. In: *AMTA* (1996), 206-211.

EAMT Workshop 1996. EAMT Machine Translation Workshop... Vienna, Austria, 29-30 August 1996. Proceedings. [Geneva: EAMT].

EAMT Workshop 1997. Language technology in your organization? 1997 EAMT Workshop, Copenhagen, 21-22 May 1997. Proceedings. [Geneva: EAMT]

Elliston, J.S.G. 1979. Computer-aided translation - a business viewpoint. Translating and the Computer: proceedings of a seminar, London, 14th November 1978; ed. B.M.Snell (Amsterdam: North-Holland), 149-158.

Flanagan, M. 1996. Two years online: experiences, challenges, and trends. In: *AMTA* (1996), 192-197.

Frederking, R. E., Brown, R. D. and Hogan, C. 1998. The DIPLOMAT rapid-deployment speech MT system. In: *MT Summit* (1997), 261-262.

Heyn, M. 1996. Integrating machine translation into translation memory systems. In: EAMT Workshop (1996), 111-124.

Humphreys, L. 1996. Use of linguistic resources like translation memories in machine translation systems. In: EAMT Workshop (1996), 101-110.

Hutchins, W. J. 1986. *Machine Translation: Past, Present, Future*. Chichester: Ellis Horwood.

Hutchins, W. J. 1993. Latest developments in machine translation technology: beginning a new era in MT research. In: *MT Summit* (1993), 11-34.

Hutchins, W.J. 1994. Research methods and system designs in machine translation: a ten-year review, 1984-1994. Machine Translation: Ten Years On. Proceedings of the second international conference...held at Cranfield University, England, 12-14November 1994...ed. Douglas Clarke and Alfred Vella. Cranfield: Cranfield University, 1998. Ch.4.

Hutchins, W.J. 1996. The state of machine translation in Europe. In: *AMTA* (1996), 198-205.

Krauwer, S. et al. 1997, eds.. Spoken language translation. Proceedings of a Workshop sponsored by the Association of Computational Linguistics and by the European Network in Language and Speech, 11 July 1997... Madrid, Spain. [Somerset, N.J.: ACL].

Lee, A. 1994. Controlled English with and without machine translation. *Aslib Proceedings* 46(5), 131-133.

Leon, M. and Aymerich 1997. PAHO systems: SPANAM and ENGSPAN. In: *MT Summit* (1997), 259-260.

Lewis, T. 1997. MT as a commercial service: three case studies. In: *MT Summit* (1997), 219-223.

McCarley, J. S. and Roukos, S. 1998. Fast document translation for cross-language information retrieval. In: *AMTA* (1998), 150-157.

McLaughlin, S. and Schwall, U. 1998. Spicing up the information soup: machine translation and the Internet. In: *AMTA* (1998), 384-397.

Mitamura, T. and Nyberg. E.H. 1995. Controlled English for knowledge-based MT: experience with the KANT system. TMI-95: Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation, July 5-7, 1995, Leuven. Belgium; 158-172.

MT Summit 1993. MT Summit IV: international cooperation for global communication, July 20-22, 1993, Kobe, Japan. [Tokyo: AAMT.]

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

MT Summit 1997. MT Summit VI: past, present, future. 29 October- 1 November 1997, San Diego, California. Proceedings, ed. By V.Teller and B.Sundheim. Washington, D.C.: Association for Machine Translation in the Americas.

Oard, D.W. 1998. A comparative study of query and document translation for cross-language information retrieval. In: *AMTA* 1998, 472-483.

O'Neill-Brown, P. 1996. JICST Japanese-English machine translation system. In: *AMTA* 1996, 257-260.

Ørsnes, B., Music, B. and Maegaard, B. 1996. PaTrans - a patent translation system. Coling 96. Proceedings... Copenhagen August 1996, 1115-1118.

Pedtke, T. R. 1997. U.S. government support and use of machine translation: current status. In: *MT Summit* 1997, 3-13.

Schaeler, R. 1996. Machine translation, translation memories and the phrasal lexicon: the localisation perspective. In: EAMT Workshop (1996), 21-33.
Schütz, J. 1996. European research and development in machine translation. *MT News International* 15 (October 1996), 8-11.
Senez, D. 1996. Machine translation feasibility study at the European Commission. In: EAMT Workshop (1996), 63-70.
Theologitis, D. 1997. EURAMIS, the platform of the EC translator. In: EAMT Workshop (1997), 17-32.
Toole, J.. Turcato, D., Popowich, F., Fass, D. and McFetridge, P. 1998. Time-constrained machine translation. In: *AMTA* (1998), 103-112.
Van der Steen, G. and Dijenborgh, B.J. 1992. Online collection and translation of industrial texts. Translating and the Computer 14... 10-11 November 1992, London. (London: Aslib), 135-164.
Vasconcellos, M. 1993. The present state of machine translation usage technology; or: How do I use thee? Let me count the ways. In: *MT Summit* (1993), 47-62.
Yang, J. and Lange, E. 1998. SYSTRAN on AltaVista: a user study on real-time machine translation on the Internet. In: *AMTA* (1998), 275-285.

JOHN HUTCHINS
University of East Anglia
Norwich NR4 7TJ, England
E-mail: WJHutchins@compuserve.com

INTERNATIONAL JOURNAL OF TRANSLATION
Vol. 15, No. 1, Jan.-June 2003

Relevance of Parallel Corpora to the Latest Developments of Machine Translation and Computer-Assisted Translation

FEDERICO GASPARI
Centre for Computational Linguistics

1. INTRODUCTION

MACHINE TRANSLATION AND COMPUTER-ASSISTED TRANSLATION

At the outset of this discussion on the practical impact of parallel corpora on machine translation (MT) systems and computer-assisted translation (CAT) tools a clarification is necessary to avoid confusion or ambiguity: machine translation is solely understood here in the strict sense of the fully automatic process in which the computer actually does the translating, and at most humans are entrusted with the supervision of what it produces (e.g. giving feedback to designers and software engineers to enhance the robustness of the MT system, updating the internal dictionaries, lexicons and terminological components), or with minor interventions taking place before, during or after the automated translation task performed by the machine.

These procedures, such as pre-editing of the source text, interactive use of MT systems and post-editing of the raw output, are typically aimed at maximising the quality of the resulting text, so as to guarantee its readability in the target language, thus avoiding serious hindrances to general understanding. In summary, machine translation is here intended as a process or activity which is automated to such an extent that it almost completely relies on the computer's performance, and accordingly the role of humans is heavily dependent on the product of the system. In this case the autonomous activity of the computer affects to a predominant extent the final translation.

As a result, in the interest of clarity in this discussion the notion of machine translation will be kept clearly distinct from that of computer-assisted translation, since the latter includes a wide range of tools that

offer support to human translators, who can avail themselves of these resources to work in a semi-automated environment. By integrating CAT tools into their working routine, humans nevertheless continue to play the leading role, while they are helped by the computer to avoid some of the