

[Written for *HLTCentral*, January 2002]

The state of machine translation in Europe and future prospects

John Hutchins

The aim of using computers for translation is not to emulate or rival human translation but to produce rough translations which can serve as drafts for published translations, as gists for information gathering, and as cross-language communication aids. The field of machine translation (MT) covers the usage, research and development of computer aids and systems ranging from production systems for large corporations to Internet aids for individuals in their own homes.

1. The recent growth of MT

The traditional use of MT is the production of translations of technical documentation, e.g. for multinational companies. The system produces 'raw' output of variable quality which has then to be revised (post-edited) by translators. Post-editing can be expensive, and a successful cost-effective option is the pre-editing of input texts (typically with a controlled 'regularized' language) to minimize incorrect MT output and reduce editing processes. An important development of this usage, now expanding rapidly (with millions of translated pages every year), is the integration of translation with technical authoring, printing and publishing processes.

Although MT software for personal computers began to appear in the early 1980s, sales were relatively low until the mid 1990s. There are now estimated to be some 1000 different MT packages on sale (when each language pair is counted separately.) Quality is not good enough for professional translators, but it is found adequate for individual 'occasional' users, e.g. for gists of foreign texts in their own language, or for communicating with others in other languages. The quality may be poor but the demand is great.

Professional translators, translation agencies and smaller companies prefer computer-based translation tools, and in particular translator workstations, often referred to by their most distinctive component as 'translation memory' systems. The most widely used originate from Europe: Trados' Translation Workbench, IBM's TranslationManager, STAR's Transit, Atril's Déjà Vu, and Eurolang's Optimizer. Each offer similar ranges of facilities and functions: multilingual split-screen word processing, terminology recognition, retrieval and management, creation and use of translation memories (bilingual text corpora of previous translations and their originals), and support for all European and many Asian languages, both as source and target languages. Finally, and not least, workstations provide access to fully automatic translation if and when required.

The Internet has produced a rapidly growing demand for real-time on-line translation. The need is for fast acquisition of foreign-language information, and top quality output is not essential. Many PC-based systems are marketed for the translation of Web pages and of

electronic mail, and there is great and increasing usage of MT services (often free), such as the well-known 'Babelfish' on AltaVista.

At the same time, the Internet is providing the means for more rapid delivery of quality translations to individuals and to small companies, and a number of MT system vendors now offer translation services, usually 'adding value' by human post-editing.

2. MT in Europe

Most of the PC-based MT software originates from Japan and the United States, and sales have been lower in Europe. However, there are notable European products: the Compendium and T1 systems (Sail Labs), Personal Translator PT (Linguattec), the iTranslator series (originally Lernout & Hauspie, now Mendez), the Reverso systems (Softissimo), the range of ProMT systems (for Russian to/from English and German); and the PARS systems for Russian and Ukrainian to and from English. Most of these systems are available in different versions for large enterprises, for independent professional translators, and for occasional (home) use, e.g. for translating Web pages and emails.

Other PC-based systems from Europe include: PeTra for translating between Italian and English; the Al-Nakil system for Arabic, French and English; the Winger system for Danish-English, French-English and English-Spanish; and the TranSmart system for Finnish-English from Kielikone Ltd.

Custom-built MT, whether for company-internal use or for clients, is a distinctive feature of European activity. Both Winger and TranSmart were initially built for particular customers; the PaTrans system was developed specifically for LingTech A/S to translate English patents into Danish. Major providers of custom-built systems are Lant n.v. and Cap Volmac Lingware Services, both specializing in controlled-language systems.

In Europe, the main users of MT systems are large translation services and multinational companies, e.g. the software group SAP is translating some 8 millions words a year; and the European Commission has seen a rapid growth in the use of Systran, now some 200,000 pages a year – mainly by non-linguist staff wanting translations for information purposes or drafts for writing documents in non-native languages. The Commission's Translation Service has developed its own workstation, EURAMIS, optimizing use of its linguistic resources, the Eurodicautom and CELEX databases, and collections of translated European Union documents (as a 'translation memory'), and providing easy access to its own MT system and other MT services.

The most distinctive feature of the European scene is the growth of companies providing software localisation (many based in Ireland), which have acquired considerable experience in the use of translation aids and MT systems, often in combination. A forum for the interchange of experience was set up in 1990, the Localisation Industry Standards Association; and Ireland has its own Software Localisation Group and a Localisation Resources Centre.

3. MT research

There is much interest in exploring new techniques in neural networks, parallel processing, and particularly in corpus-based approaches: statistical text analysis (alignment, etc.), example-based machine translation, hybrid systems combining traditional linguistic rules

and statistical methods, and so forth. Above all, the crucial problem of lexicon acquisition (always a bottleneck for MT) is receiving major attention by many academic research groups, and the large lexical and text resources (e.g. from the LDC and ELRA) are being widely and fruitfully exploited.

The most innovative area of current research is automatic translation of spoken language, a distant vision until developments in speech technology in the 1980s. The main centres are ATR in Japan, the Carnegie-Mellon University (USA), the University of Karlsruhe (Germany), all collaborating in a project (C-STAR consortium) to develop speaker-independent real-time telephone translation systems for Japanese, English and German – initially for hotel reservation and conference registration transactions. Also in Germany is the government-funded Verbmobil project to develop a portable aid for business negotiations (German, Japanese, English), and involving numerous German university groups in fundamental research on dialogue linguistics, speech recognition and MT design (see: www.dfki.uni-sb.de/verbmobil/). Speech translation attracts much publicity, but few observers expect dramatic developments in the near future. However, in the meantime, many marketed MT systems include voice input and output – i.e. speech-to-text and text-to-speech conversion upon a text-to-text base.

The planned accession of states in Central and Eastern Europe to the European Union has stimulated research on MT and translation tools for languages such as Czech, Polish, Hungarian, Slovenian, Estonian and Bulgarian – not just for supporting translation of treaty and other legal documents but also for enabling public access to information resources. Indeed, today many projects funded by the European Union within the broad field of human language technology (www.cordis.lu/ist/) involve multilingual tools of all kinds and include translation aids, usually within a restricted subject field and often in controlled conditions, and many are designed for Internet applications.

Mention should also be made of research on systems for ‘minority’ languages in Europe, such as Basque, Catalan and Galician in Spain and immigrant languages such as Hindi, Bengali and Gujarati in the United Kingdom. The need is both for full translation systems and for translation aids, dictionaries, glossaries, bilingual corpora of authorized translations, etc.

Finally, the Internet has demonstrated an urgent need to replace the existing systems, developed for well-written scientific and technical documents and assuming human post-editing, by systems and translation aids which are developed specifically to deal with the kind of colloquial (often ill formed and badly spelled) messages found in emails and chatrooms, where there is no possibility of any human revision. The old linguistics rule-based approaches are probably not equal to the task on their own, and we may expect corpus-based methods making use of the voluminous data available on the Internet itself to form the basis of future systems for this application.

4. MT and future translation demand

One impact of the Internet may well concern the future nature of the software itself. What users of Internet services are seeking is information, in whatever language it may have been written or stored – translation is just a means to that end. Users will want seamless integration of information retrieval, extraction and summarization systems with translation. Research has begun in such areas as cross-lingual information retrieval, multilingual summarization,

multilingual text generation from databases, and so forth – in Europe and elsewhere – and before many years there may well be systems available on the market and the Internet.

Perhaps in future years there will be fewer ‘pure’ MT systems (commercial, on-line, or otherwise) and many more computer-based tools and applications where automatic translation is just one component. Integrated translation software will be the norm not only for the multinational companies but also available and accessible for anyone from their own computer (whether desktop, laptop, or network-based, etc.) and from any device (television, mobile telephone, etc.) interfacing with computer networks. It will not spell the end of the ‘pure’ MT system completely, but be a demand-led expansion of the provision of translation software more accessible and usable in the ‘information society’.

Where translation has to be of publishable quality, both human translation and MT have their roles. Machine translation is demonstrably cost-effective for large scale and/or rapid translation of (boring) technical documentation, (highly repetitive) software localization manuals, and many other situations where the costs of MT plus essential human preparation and revision or the costs of using computerized translation tools (workstations, etc.) are significantly less than those of traditional human translation with no computer aids. By contrast, the human translator is (and will remain) unrivalled for non-repetitive linguistically sophisticated texts (e.g. in literature and law), and even for one-off texts in specific highly specialized technical subjects.

In addition, it is probable that the ready availability of low-quality MT output from Internet services and from commercial software will create a demand for high-quality human translations from people who have previously had no exposure to translation facilities.

For the translation of texts where the quality of output is much less important, machine translation is often an ideal or even the only solution. For example, to produce translations of scientific and technical documents that may be read by only one person who wants to merely find out general background information and/or specific data, MT will increasingly be the only answer.

For business correspondence, there will probably always be a role for the human translator, 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, MT is already the only feasible solution.

As for spoken language translation, there can be no prospect of automatic translation replacing the interpreter of diplomatic exchanges. While we can envisage MT of speech in highly constrained domains (e.g. telephone enquiries, banking transactions, computer input) it seems unlikely that automatic speech translation will extend to open-ended interpersonal communication.

In the past there has been tension between the translation profession and those who advocate and research computer-based translation tools. But now it is apparent that MT and human translation can and will co-exist in relative harmony. In many cases, 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; the real-time translation of television subtitles; the translation of information from databases; the on-line translation of Web pages, etc.

We may expect more such new applications as global communication networks expand and as the usability and usefulness of the less-than-perfect output of MT systems are recognized

by a wider public. In this broader context, it is becoming more appropriate to perceive this area of human language technology as concerned with bilingual (multilingual, or translingual) communication aids rather than with 'translation' systems as such.

5. Further information.

The website of the European Association for Machine Translation (www.eamt.org) includes information about MT (including details of commercial systems in the *Compendium of translation software*) and links to other resources worldwide.

W. John Hutchins (b.1939), university librarian by profession, is the author of articles and books on linguistics, information retrieval, and in particular machine translation, mainly surveys and historical works – many available from his website. Principal works: *Machine translation: past, present, future* (Chichester: Ellis Horwood, 1986); *An introduction to machine translation* [with Harold Somers] (London: Academic Press, 1992); editor of *MT News International* (1991-1997), compiler of *Compendium of translation software* (2000); editor of *Early years in machine translation: memoirs and biographies of pioneers* (Amsterdam: John Benjamins, 2000). He is active in the European Association for Machine Translation (president since 1995) and the International Association for Machine Translation (president, 1999-2001).

Email: WJHutchins@compuserve.com; Web: <http://ourworld.compuserve.com/homepages/WJHutchins>.