# A Brief Overview of Machine Translation in Europe and North America

**John Hutchins**

**Abstract:** 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 means for accessing foreign-language information, 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. It is a major area of research in both Europe and North America.

**Keywords**: Machine translation, Europe, America

### Types of MT systems

MT systems are widely available in Europe and North America in three basic versions: 'corporate' or 'enterprise' versions for large companies; 'professional' versions for independent professional translators; and 'home' or 'personal' systems for occasional users, e.g. for translating Web pages and emails. The 'home' systems are the most basic type, consisting of little more than the core translation engine. The corporate systems include many additional aids, for pre- and post-editing of texts, for terminology management, for project control, etc. The 'professional' systems provide a selection of those facilities found to be most suitable for translators.

Although there have been improvements since MT began in the 1950s, it is true to say that translation quality is still not good enough for publication without revision; Human intervention or interaction is necessary for publication-quality translation. Hence, in many environments, the need for editing facilities and for the 'control' of text input (e.g. use of regularized language). On the other hand, the quality of MT output is often found to be adequate for less demanding uses, e.g. for individual 'occasional' translation, for identifying the main content of foreign texts or for non-technical communication between languages.

MT software is available from a large number of European and North American vendors and covering virtually all European language pairs. Here we can mention only the most notable (for a full listing see the *Compendium of translation software* at http://www.hutchinsweb.me.uk/ Compendium.htm). Nearly all cover the major European languages (English, French, German, Italian, Spanish), and many of them also translate from less common Languages (Greek, Polish, Russian, Hungarian, Turkish, etc.) and from and into Arabic, Chinese, Japanese, Korean, etc. In addition, there are many systems specifically designed for particular language pairs: English-German (Personal Translator PT), English-Italian (PeTra), English-Finnish (TranSmart), Hungarian-English (Morphologic), Arabic-English (Al-Mutarjim Al-Arabey, Al-Nakil, Al-Wafi); French-German (FB-Active), German-Russian (PROMT), Russian-Ukrainian (PARS), Portuguese-Spanish and other languages (Falatudo), Catalan-Spanish (interNOSTRUM), etc.

### MT for special purposes

Apart from commercial systems there continue to be custom-built systems for company-internal use or for corporate clients. In the United States, the PAHO (Pan American Health Organization) developed on-site systems for English and Spanish in the early 1980s, followed later by English-Portuguese; the Smart Corporation continues to develop customized systems for most European languages for large corporate clients; and European providers of custom-built systems include ESTeam specializing in controlled-language systems.

Many large translation services and multinational companies use MT systems for translating large volumes of texts, e.g. in the United States government institutions (DARPA, USAF, etc.) and large corporations (Xerox, Ford, General Motors, etc.). Major users in Europe are companies such as SAP and Siemens, and in particular the European Commission.

One of the most distinctive features of the European scene are translation companies providing localisation of documentation and products – these companies have acquired considerable experience in the use of translation aids and MT systems. Related to this activity is the development of software for the localisation of websites. With the growth of the Internet, many companies offer information about their products and services, which increasingly needs to be made available in other languages. The information has to be updated regularly, and software such as IBM Websphere has been developed specifically for translating webpages as and when required.

Automatic translation of electronic mail has been relatively neglected. Most PC based MT software is designed for this usage, but it is clear that more specialised software is desirable. To meet this need, the Translution company is marketing customized

software for translation of company-internal email communication.

In contrast to the situation in Japan and other Asian countries, the application of MT to patents has been relatively neglected. There are only three systems specifically for translating patents: the PaTrans and SpaTrans systems developed for LingTech A/S to translate English patents into Danish; and the APTrans system designed for generating multilingual patent claims from controlled English language input.

**MT and the Internet**

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 at all essential. Many MT systems are marketed for the translation of Web pages and of electronic mail, and there is great and increasing usage of MT services (many free), such as the well-known 'Babelfish' on AltaVista – and now also available on Yahoo. Others include FreeTranslation, Google Translator, Tarjim, WorldLingo, and many more online services are being added (see the *Compendium of translation software*), both for specific language pairs and for the 'major' languages (English, French, German, Spanish, Arabic, Japanese, Korean, Chinese).

**MT research**

Until the mid 1990s, most MT research was still based on the implementation of lexical and grammar rules (with translation via an interlingua or at least 'deep structure' representations) in what is now called rule-based machine translation (RBMT). Currently, the dominant paradigms of MT research are corpus-based. In statistical machine translation (SMT), words and 'phrases' (sequences of two or three words) from a bilingual corpus (of original texts and their translations) are aligned as the basis for a 'translation model' of word-word (and phrase-phrase) frequencies. Translation involves the selection of the most probable words in the target language for each input word and the determination of the most probable sequence of the selected words (on the basis of a monolingual 'language model'). Example-based machine translation (EBMT) involves similar alignment of bilingual data, but here the translation units are larger than individual words or short word sequences; input sentences are matched against phrases or clauses (examples) in the corpus, then equivalent phrases in the target language are extracted, and adapted and combined in acceptable output sentences. Both methods make substantial use of large bilingual corpora, but where SMT is based exclusively on statistical correlations, EBMT applies both statistical techniques and linguistics-based methods similar to those of earlier RBMT approaches.

Although SMT research now dominates MT research, the great majority of commercial systems are RBMT systems. Few SMT systems have reached public operational status. The leader has been Language Weaver offering translation systems for Arabic, Chinese, French, German, Persian, Romanian, Spanish, etc. to and from English. Most recently, the online 'Google Translate' service has began offering its own internally-developed SMT system for Arabic, Chinese, Japanese and Korean into English – using the resources of Google's massive text databases.

One of the advantages of the SMT approach is that systems can be developed quickly – significantly quicker than RBMT systems. There are many examples in the MT literature of recent years (see the *Machine Translation Archive*.) One example is the statistics-based SpaTrans system which has been developed for English-Danish translation of patents by LingTech A/S in Denmark. It has been evaluated in comparison with the rule-based PaTrans system developed in the 1990s from the Eurotra model. On the whole the output quality of the SMT system compares well with the rule-based system – relatively simple adjustments can be envisaged to improve treatment of word order and noun-verb agreement. However, a major problem is the occurrence of new terminology, resulting in many 'unknown' words in the output. It is, of course, a problem for all MT systems. In the case of SMT, there are two possible solutions: the addition of a bilingual dictionary database, or the addition of more domain-specific texts to the bilingual corpus. The latter is generally preferred as it causes less disruption to the core statistical model.

Probably the most significant development in MT research in Europe is the establishment of the Euromatrix project (based at Edinburgh University). Its aim is the development of open-source MT technologies applicable to all language pairs within Europe, based on hybrid designs combining statistical and rule-based methods. There will be a particular emphasis on languages of new member states of the European Union, and on systems for translating technical, social and legal documentation.

The most innovative area of current research is automatic translation of spoken language. 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. Until recently, there was also in Germany the government-funded Verbmobil project to develop a portable aid for business negotiations (German, Japanese, English). Speech translation attracts

much publicity, but few observers expect dramatic developments in the near future. 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.

The 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. 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.

**MT and human translators**

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.

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 – many developed initially by European companies. The most widely used currently are: SDL, Transit, Déjà Vu, MultiTrans, LogiTerm, Wordfast, and ProMemoria. 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.

Significant ‘by-products’ of the corpus-based MT research have been further development of aids for translators, not just improvements in translation memories, their creation and exploitation, but also systems for error detection and correction and for automatic text prediction, i.e. suggestions for text completion to aid human translators who frequently translate similar technical documents.

**MT in the future**

For the translation of texts where the quality of output is not important, machine translation is often the only solution. For example, it will be uneconomic to produce human translations of scientific and technical documents just for general background information and/or specific data. In these cases, MT will increasingly be the only answer. And there are new applications 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; the translation of electronic mail; etc.

The Internet will drive changes in the nature and application of MT. 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 automatic translation. There is now active research in cross-lingual information retrieval, multilingual summarization, multilingual text generation from databases, and so forth.

Existing systems have been developed for well-written scientific and technical documents and assume human post-editing. Internet usage demands systems specifically for the kind of colloquial (often ill formed and badly spelled) messages found in emails and chat rooms. The old linguistics rule-based (RBMT) 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 as the basis of future systems for this application.

---

**W. John Hutchins** is the author of articles and books on linguistics, information retrieval, and in particular machine translation - many available from his website (http://www.hutchinsweb.me.uk). He is active in the European Association for Machine Translation (president 1995-2004) and the International Association for Machine Translation (president, 1999-2001).

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* (now on his website) (2000 to the present) and of the *Machine Translation Archive* (http://www.mt-archive.info) (2004 to the present); Editor of *Early years in Machine Translation: Memoirs and Biographies of Pioneers* (Amsterdam: John Benjamins, 2000).