Commercial systems

The state of the art

John Hutchins
University of East Anglia, Norwich

1. Introduction

In a general overview of the availability and potential usefulness of commercial machine translation (MT) systems and translation tools, it is important to distinguish three basic types of translation demand: dissemination, assimilation, and interchange.

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 use of MT for **dissemination** purposes 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 these purposes has been joined by developments in translation tools (e.g. terminology databases and translation memories – TMs), often integrated with authoring and publishing processes. These 'translation workstations' are more attractive to human translators. Whereas with MT systems they 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 the translators are in control of computer-based facilities producing output which they can accept or reject as they wish.

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 or database resource – and generally, as quickly as possible. The use of MT for **assimilation** has been met as, in effect, a by-product of systems designed originally for the dissemination application, since some users found that they could extract what they needed to know (e.g. for screening and/or information gathering) from the unedited MT 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 (often known as "gisting") has undoubtedly grown substantially.

Related to this application is translation within multilingual systems of information retrieval, information extraction, database access, etc. Here MT systems operate as components of **information access** systems, i.e. translation software is integrated in other systems: (a) 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; (b) systems for extracting information (e.g. product details) from texts, in particular from newspaper reports; (c) systems for summarising texts; and (d) systems for interrogating non-textual databases. As yet, however, there are few commercial systems available in this area.

The third type of demand is that for translation between participants in one-to-one communication (telephone or written correspondence). In this **interchange** use, the situation is changing quickly. The demand for translations of electronic texts on the Internet, such as electronic mail and discussion groups, is developing rapidly. In this context, 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 here, since they can operate virtually or in fact in real-time and on-line and there is 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.

2. Types of systems

At the present time we may distinguish the following types of systems and their most appropriate areas of application:

a) mainframe, workstation and/or client-server systems on intranets of large organisations;

b) MT systems for professional translators;

c) translator's workstations for professional translators operating either on company intranets or independently;

d) computerised translation tools: dictionaries, terminology management software, TM systems;

e) MT systems for occasional users and/or casual home use;

f) systems designed for Internet use and/or for translating Web pages, either for company or individual use;

g) MT services on the Internet providing translations on demand.

Traditionally, MT systems have been divided according to architectures: direct translation, transfer-type, interlingua-based, statistics-based, example-based, etc., but these distinctions are largely irrelevant to and hidden from users, and they are ignored in this paper. In general, however, it may be pointed out that commercial systems are based usually on well-tested approaches – for obvious reasons – and these tend to be based on the older traditional (linguistics rule-based) strategies developed from the 1960s to the late 1980s. More recent developments in MT research in the 1990s based on text corpora – the statistics-based and example-based approaches – have not yet had much impact on the commercial MT scene. Increasingly, there are however systems incorporating example-based methods, and of course the translator's workstations make considerable use of statistics-based facilities for the creation and utilisation of TMs, i.e. bilingual corpora of previous translations and their originals.

As throughout the computing industry, there has been a *de facto* standardisation of hardware, operating systems and inter-compatibility. In particular, for the smaller systems, the standards are PC compatibles, Pentium CPUs, Microsoft Windows 95, 98, ME, 2000,

NT, etc. A few are still available for Microsoft DOS systems, and some (although increasingly rarely) are designed for Macintosh equipment. As for Internet access, nearly all systems either include or run with Netscape *Navigator*, Netscape *Communicator*, or Microsoft *Internet Explorer*. MT products for Japanese, Chinese, and Korean generally require additional software (e.g. *Japanese Windows*, *Japanese Language Kit*), and occasionally run only on proprietary hardware.

The focus of this chapter will be the development and use of commercially available systems for dissemination, i.e. for aiding the production of "publishable" quality translations. Other applications will be treated more briefly. Since changes in the MT market are so rapid: every year there are many new systems, many developments in old systems (new platforms, new languages, etc.), companies merge or cease trading, and many products become no longer available. Full details of systems available – including those mentioned here for illustrative purposes – may be found in the *Compendium of translation software*. This is a regularly updated listing of current commercial MT systems and computer-aided translation support tools (including translator's workstations, terminology management systems, electronic dictionaries, localisation support tools, etc.)

2.1. Mainframe, client-server and workstation systems

The oldest MT systems are those developed originally for mainframe computers, e.g. the *Systran*, *Logos* and Fujitsu (*Atlas*) systems. *Systran*, 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 West European languages (French, German, Italian, Spanish, Portuguese), Japanese, Korean, etc. Likewise, *Logos*, originally marketed for German to English, was later 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).

Mainframe systems – much improved from their earlier 1960s and 1970s designs – are still available, primarily now it appears for large companies or organisations wanting to include a MT engine in already existing documentation systems, but for most purposes, large-scale systems take the form of workstations or client-server systems operating over company intranets. A popular choice for the workstation and/or server has been the Sun SPARCstation, and many of the older larger systems are still available for this platform. However, some Japanese computer companies chose to develop MT software for their own equipment, and some are still available commercially only on their proprietary platforms.

Needless to say, the prices of client-server systems make them affordable only for large companies or organisations with large translation services. From the mid 1990s onwards, most of these systems have begun to appear in cheaper versions for personal computers – although often with substantially smaller dictionary resources and without facilities for working in groups and networks.

The main customers and users of mainframe and client-server systems are the multinational companies exporting products and goods in the global market. The need is primarily for translation of promotional and technical documentation. 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. 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. Translation systems are now being seamlessly integrated with other computer-based systems

for the support of technical writers, not just assistance with terminology, but also on-line style manuals and grammar aids.

There are numerous examples of the successful and long-term use of MT systems by multinationals for technical documentation. One of the best known has been the application of the *Logos* system at the Lexi-Tech company in New Brunswick, Canada; initially for the translation into French of manuals for the maintenance of naval frigates, later as a service for many other large translation projects. *Systran* has had many large clients: Ford, General Motors, Aérospatiale, Berlitz, Xerox, etc. Users of *Logos* have included Ericsson, Osram, Océ Technologies, SAP and Corel. The *Metal* German-English system (no longer on the market) 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.), and that the user has available (or has the resources required to acquire or to create) a terminological database for the particular application. The creation of terminology databases and the maintenance of large dictionaries demands considerable initial and continuing expenditure, which can usually be justified only if translation production is on a large scale. In fact, it is always desirable for company documentation to be consistent in the use of terminology. In addition, many companies 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.

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 customisation may embrace 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. A further step is the implementation of a company-specific controlled language, not just for standardisation but for reducing well-known problems of MT such as lexical and structural ambiguities in source texts. The amount of work involved in such customisation and in the pre-editing control of input may not be justifiable unless output is in a number of different languages.

Large savings are reported by many companies that have installed MT systems, in some cases there have been reductions in the costs of producing finished translations of up to 40 or 50%, and nearly all companies report much faster throughputs. However, it must be stressed that it is only the larger organisations dealing with 100,000 pages a year or more that can expect such dramatic savings. Smaller companies and translation services may gain only in terms of speed of production and may experience few cost savings. The situation is, however, changing rapidly, and cheaper more powerful MT systems, combined with cheaper and more powerful publishing and authoring systems, will probably bring comparable savings to a wider range of companies and services.

Multinational companies at many locations in different countries are often linked by internal networks (intranets). In this environment, translation jobs can be passed easily in electronic form from one office or branch of the organisation to another. Indeed, a document may be authored in one location, sent for translation at another, and printed and distributed at a third. There are a number of client-server systems on the market, e.g. *Atlas* (from Fujitsu), *Systran Enterprise*, *Enterprise Translation Server* (SDL*)*, *TranSphere* (AppTek), and *TranSmart* (Kielikone), and there are also companies that develop client-server software for specific customers, nearly always large government organisations or multinational

corporations. The oldest is Smart Communications Inc. of New York, which has built systems for Ford, Citicorp, Canadian Department of Employment and Immigration, etc.; others include ESTeam Inc. of Greece, and LANT n.v. of Belgium

2.2 Translator's workstations

In the 1990s, the options for large-scale computer-based translation production broadened with the appearance on the market of translator's workstations. These combine multilingual word processing, means of receiving and sending electronic documents, facilities for document scanning by OCR (optical character recognition), terminology management software, facilities for concordancing, and in particular TMs. The latter facility enables translators to store original texts and their translated versions side by side, 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 TM 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 the TM, 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. Translator's workstations also give access to terminology databases, in particular to company-specific terminology, for words or phrases not found in the TM. In addition, many translator's workstations are now offering full automatic translations using commercial MT systems. 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.

The translator's workstation has revolutionised the use of computers by translators. Translators 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 staff resources. TMs 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.

2.3 Localisation support tools

One of the fastest growing areas for the use of computers in translation is software localisation. Here the demand is for producing documentation in many languages to be available 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, TMs in translator's workstations are the obvious solution. Among the first in this field was the large software company SAP AG in Germany, using older MT systems, *Metal* and *Logos*. Most localisation, however, is based on the TM and workstation approach – mainly *Transit*, *Déjà Vu*, and the *Trados Workbench*.

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. The need for a general translation and management support environment for a wide variety of TM, MT and other productivity tools is seen as fundamental, and a number of companies are producing "localisation support tools", for managing and routing localisation among translators, software engineers, project managers, for efficient use of different tools during overall processes, for automated updating, unified file tracking, etc.

The translation and management requirements of software localisation have been sufficiently distinct for the creation of a dedicated organisation (Localisation Industry

Standards Association, LISA), which holds regular seminars and conferences throughout the world.

2.4 Systems for independent professional translators

For the independent translator, the translator's workstation may be no more affordable than the larger MT systems. Professional translators not employed by large organisations have currently two options: (a) relatively powerful systems capable of running on widely available computer equipment, e.g. *Windows*-based PC systems, and (b) translation support tools, such as terminology management systems and TM programs.

Most of the vendors of client-server systems also have systems on the market designed primarily for the demands of the professional translator user, i.e. systems that have facilities for post-editing and publishing, and that can be used with terminology databases and sophisticated word processing facilities. In origin, these systems are either downsized versions of mainframe (or client-server) systems or enhanced versions of cheaper PC systems. In the case of the former, often the same range of languages is covered as for the larger intranet versions, e.g. *Systran Professional*, and the two systems from the Pan American Health Organization (*Spanam* Spanish to English, and *Engspan* English to Spanish). What these systems lack in comparison with the intranet client-server systems are generally the wide range of document formatting and conversion facilities and sometimes the complete range of text processing compatibility. However, even this situation is changing as standalone computers become more powerful, and as users' demands become clearer, so that increasingly these "professional" systems for the independent translator are acquiring the range of facilities found previously only in the largest mainframe and client-server systems.

2.5 Translation support tools

Just as large companies may well prefer translation workstations to fully-fledged MT systems, the individual professional translator may not want to purchase a MT system that may cover only some of the languages required. Since the mid 1980s there has been a wide range of translation aids, some designed originally for workstations in larger organisations, intended primarily for individual translators for use on PC-type equipment.

Electronic dictionaries (usually in CD-ROM form) are available from nearly all dictionary publishers, and from many companies supplying computer software. There are also many dictionaries accessible on the Internet.

Terminology management software provides facilities for professional translators to create, update and revise their own lexical resources, whatever the languages concerned. Typical facilities include means for downloading from on-line or other electronic databases. Software for TMs in individual packages (as opposed to components of translator's workstations) is being marketed by a number of vendors. These programs allow individual professional translators to build their own stores of searchable and aligned bilingual databases of original texts and their translations. Most can cope with texts both in any language written in Roman characters, and some with non-Roman scripts.

2.6 Systems for non-professional (home) users

The basic need of the non-professional user of translation software is primarily as a means of access to foreign language texts, to find out what a text in an unfamiliar or unknown language is about. What matters is the message. It is usually not essential to have a "perfect" translation. Any of the systems already mentioned can serve this need; indeed in earlier years one of the main uses of mainframe MT systems was the provision of rough translations, i.e. the unedited crude output, for the purposes of intelligence analysis or for scientific and technological reviews. At the European Commission, one of the principal uses

of the *Systran* system is still the production of crude (sometimes lightly edited) translations for rapid surveys of documentation.

Software for personal computers began to appear in the early 1980s in systems from ALPS and from Weidner. Their output was at a level of quality suitable only for information assimilation use, but they were too expensive for the casual home user. In fact they were bought mainly by professional translators, who found them frustratingly unsuited for their needs. This experience may have convinced professional translators that PC translation software would always be useless for their purposes, but the more recent "professional" systems described above are changing this perception.

It was not until PC equipment and software were much reduced in price during the early 1990s that this large potential "non-professional" market was opened up. Earliest in the field were the Japanese computer companies, selling systems, usually for English to and from Japanese, and designed to run on their own microcomputers. In the United States the earliest vendors were Linguistic Products with its series of *PC-Translator* systems, and Globalink, with its well-known *Power Translator*. They have been succeeded by numerous other vendors, many surviving only a few years in this very competitive market. Many of the producers of client-server systems have sold versions of their systems for the home or non-professional market, but not always with the same large range of language pairs.

Finally, it may be noted that there is a proliferation of particularly inexpensive products, marketed as "translation systems" but which in fact are little more than electronic dictionaries. They sell presumably because of the widespread belief among those unfamiliar with translation that all that is needed in order to translate something is a bilingual dictionary.

Sales of PC translation software showed a dramatic rise during the 1990s. There are now estimated to be some 1,000 different MT packages on sale (when each language pair is counted separately.) For example, in Japan one system (*Korya Eiwa*, for English-Japanese translation) was said to have sold over 100,000 copies in its first year on the market. A recent development for many home-use systems has been the addition of facilities for voice input and voice output – this is not, of course, true translation of spoken language (conversation, etc.) but speech-to-text conversion, text-text translation, and text-to-speech synthesis.

Though it is difficult to establish how much of the translation software sold in large numbers is used regularly after initial purchase (some cynics claim that only a very small proportion is tried out more than once), there is no doubting the growing demand for "occasional" translation, i.e. by people from all backgrounds wanting gists of foreign text in their own language, or wanting to communicate in writing with others in other languages, however poor the quality. It is this latent market for low-quality translation, untapped until very recently, which is now being exploited. As a consequence, many products have to be treated with caution – in fact, they may not even meet minimal standards for crude "information-only" translation.

2.7 MT for the Internet

The largest area of growth for translation demand is now undoubtedly based on use of the Internet. There is the need of the occasional user for software to translate web pages, e-mail and other Internet resources and texts, either off-line or on-line, and the availability of on-demand Internet-based translation services for companies.

There has been a rapid increase in MT software products designed specifically for online translation of web pages. Japanese companies such as Fujitsu, Toshiba, Hitachi and NEC led the way, primarily with systems for translating from English into Japanese. They

were followed quickly elsewhere, and nowadays, nearly all systems for home users incorporate web-page translation as standard features.

Equally significant has been the use of MT for e-mail and for "chat rooms", many of the online systems have facilities for this application. In addition, most home-use software is designed for this use, and some are intended specifically for e-mail and/or specifically for chat.

As well as these online systems there are now many Internet services offering translation facilities, many of them free. One of the earliest and probably still best known example is the AltaVista translation service *Babelfish*. There have been many followers, although some offer not full translations but little more than on-line bilingual or multilingual dictionaries. The latter are undoubtedly serving a real need; even the use of the AltaVista service is apparently mainly for translating individual words or short phrases. When translations from on-line services are of complete sentences, the output is often poor. None of the systems has been designed specifically for translating the kind of colloquial, jargon-filled, and often "ungrammatical" language found in e-mail and on-line discussion forums.

At the same time, however, there are now many network-based translation services for on-demand professional-level translation, generally with human revision as an option. 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. In most cases, clients have the option of receiving unedited translations or versions post-edited by the suppliers' own professional translators.

In the future, we may expect many more online translation services. There will be both a wider range of languages and a wider variety of charging methods. We may also expect to see services designed for particular domains and subject areas, since systems restricted to specific subjects have typically produced better quality output than general-purpose systems. Users of online translation systems (whether charged or free) will expect continued improvements, and this will be more likely with specialised services than with non-specialised ones.

2.8 MT for information access

The growing use of the Internet is highlighting the need for systems that combine translation with other language-oriented facilities, in particular database searching, information retrieval and summarisation. As yet, however, there are few such systems available commercially. Most of the web-page translators could be used for this purpose, although few enable search terms to be formulated and translated before searching Web resources.

It is to be expected that in future this will be one of the main growth areas. Several research projects supported by the European Union combine MT with programs for information access, information extraction, and summarisation. There is equally intensive attention to this area in North America, in Japan, Korea, China and other Asian countries. Many companies are directing their efforts to the development of products for the information marketplace.

3. Language coverage

From the very beginning of the commercialisation of MT systems, the major European languages have been well covered. Translation from English into French, German, Italian, Spanish, and Portuguese, and from these languages into English, is available from all the main vendors, and in most cases with versions for large organisations, as client-server systems, for professional translators, for home users, and for web-page and e-mail

translation. In some cases, products are dedicated to particular pairs, e.g. German and English and Spanish and English.

Systems for other European languages are less common. The Scandinavian languages are relatively poorly covered, and although Russian was the main focus of the earliest MT research, there are now fewer products for this language (and for other Slavic languages) than for western European languages. Other languages of Europe have so far been neglected by the main vendors; there are no "professional" quality translation systems for Greek, Hungarian, Rumanian, Serbian, Catalan, or any of the Celtic languages.

In the 1980s, nearly all Japanese computer companies began marketing MT systems, predominantly between English and Japanese. In recent years, many more systems have appeared, a large number specifically for Internet/web use, which are obviously meeting a great demand in Japan. The older mainframe or workstation systems are now marketed also in *Windows* versions for either English to Japanese and/or Japanese to English, and almost every month there appears a new inexpensive system for translation between these two languages. But products for translation between Japanese and English come not only from Japanese companies; there is competition also from companies of US origin.

In contrast to Japanese, there are still few Chinese to and from English systems of reasonable quality, and most systems are intended for primarily non-professional use (interactive composition). The situation is slightly better for Korean to and from English, with some good quality enterprise systems, although there are also many low-quality systems.

Other languages are even more poorly served. There have been surprisingly few systems for Arabic, despite the obvious potential market, and only one Hebrew system is marketed at present. While there have been systems for some African languages (by EPI-USE Systems (Pty) Ltd., South Africa), there are many languages still not covered by commercial systems, e.g. Indonesian, Malay, Vietnamese, Thai, and languages of the Indian sub-continent (not even Hindi, Urdu and Bengali).

In principle, most translator's workstations are designed for use with a wide range of languages; they do not need programs for linguistic analysis and synthesis, only for dealing with strings of characters and words. However, the need for greater sophistication in the alignment programs of TM systems makes them less suitable for some languages than others, particularly non-European languages. However, although designed initially and primarily for languages using Roman alphabets, workstations are increasingly available in versions suitable for use with languages such as Arabic, Chinese and Japanese.

4. Conclusion

After many years of development, commercial MT systems are now capable of serving well the demands of multilingual companies and professional translators seeking cost-effective production of good-quality translation for dissemination purposes. This is particularly the case for translation between the major languages of the global marketplace. There remain, however, many gaps for "minor" languages, including those of Eastern Europe, Africa, and India.

Systems for assimilation purposes (for the less-demanding "occasional" user) are also widely available, with good language coverage on the whole. However, these systems often give poor quality output, even for well-written source texts, let alone the low-level writing on e-mail and other Internet applications. There is clear need for improved quality in this area of commercial software, and even more for some consumer guidance in order that potential purchasers are not misled by exaggerated claims.

Further reading

The *Compendium of Translation Software* is available online at the website of the European Association for Machine Translation (www.eamt.org). Older editions of the *Compendium* – for tracking changes and developments – can be seen on the current author's website (http://ourworld.compuserve.com/homepages/WJHutchins/compendium.htm).   For   earlier surveys of MT systems see also Hutchins (1996, 1999, 2002).

References

The following articles are all available on the author's website:
        http://ourworld.compuserve.com/homepages/WJHutchins/

Hutchins, J. 1996. "'Computer-based Translation Systems and Tools". *ELRA Newsletter* 1(4), December 1996, 6-9.
Hutchins, J. 1999. "The Development and Use of Machine Translation Systems and Computer-based Translation Tools", in *International Conference on Machine Translation & Computer Language Information Processing*, 26-28 June 1999, Beijing, China. Proceedings of the conference, ed. Chen Zhaoxiong, 1-16. [Beijing: Research Center of Computer & Language Engineering, Chinese Academy of Sciences.]
Hutchins, J. 2002. "'The State of Machine Translation in Europe and Future Prospects", available on: http:// www.hltcentral.org.