Current commercial machine translation systems and computer-based translation tools: system types and their uses

John Hutchins
[Email: WJHutchins@compuserve.com]
[Website: http://ourworld.compuserve.com/homepages/WJHutchins]

1. General factors and issues

1.1. Introduction

Why should we be interested in using computers for translation at all? The first and probably most important reason is that there is just too much that needs to be translated, and that human translators cannot cope. A second reason is that on the whole technical materials are too boring for human translators, they do not like translating them, and so they look for help from computers. Thirdly, as far as large corporations are concerned, there is the major requirement that terminology is used consistently; they want terms to be translated in the same way every time. Computers are consistent, but human translators tend to seek variety; they do not like to repeat the same translation and this is no good for technical translation. A fourth reason is that the use of computer-based translation tools can increase the volume and speed of translation throughput, and companies and organizations like to have translations immediately, the next day, or even the same day... The fifth reason is that top quality human translation is not always needed. Because computers do not produce good translations, some people think that they are no use at all to anyone. The fact is that there are many different circumstances in which top quality is not essential, and in these cases, automatic translation can be and is being used widely. Lastly, companies want to reduce translation costs and, on the whole, with machine translation (MT) and translation tools they can achieve reductions. Any one of these reasons on its own can be sufficient justification for using and installing either MT systems or computer translation aids.

1.2. Brief history

Many are under the impression that MT is something quite new. In fact, it has a long history (Hutchins, 1986, 2001) – almost since before electronic digital computers existed. In 1947 when the first non-military computers were being developed, the idea of using a computer to translate was proposed. In July 1949 Warren Weaver (a director at the Rockefeller Foundation, New York) wrote an influential paper which introduced Americans to the idea of using computers for translation. From this time on, the idea spread quickly, and in fact machine translation was to become the first non-numerical application of computers. The first conference on MT took place in 1952. Just two years later, there was the first demonstration of a translation system in January 1954, and it attracted a great deal of attention in the press (Hutchins 2004). Unfortunately it was the wrong kind of attention as many

readers thought that machine translation was just around the corner and that not only would translators be out of a job but every body would be able to translate everything and anything at the touch of a button. It gave quite a false impression. However, it was not too long before the first systems were in operation, even though the quality of their output was quite poor. In 1959 a system was installed by IBM at the Foreign Technology Division of the US Air Force, and in 1963 and 1964 Georgetown University, one of the largest research projects at the time, installed systems at Euratom and at the US Atomic Energy Agency. But in 1966 there appeared a rather damning report for MT from a committee set up by most of the major sponsors of MT research in the United States (ALPAC 1966). It found that the results being produced were just too poor to justify the continuation of governmental support and it recommended the end of MT research in the USA altogether. Instead it advocated the development of computer aids for translators. Consequently, most of the US projects – the main ones in the world at that time – came to an end. The Russians, who had also started to do MT research in the mid 1950s concluded that if the Americans were not going to do it any more then they would not either, because their computers were not as powerful as the American ones. However, MT did in fact continue, and in 1970 the Systran system was installed at the US Air Force (replacing the old IBM system), and that system for Russian to English translation continues in use to this day. The year 1976 is one of the turning points for MT. In this year, the Météo system for translating weather forecasts was installed in Canada and became the first general public use of a MT system. In the same year, the European Commission decided to purchase the Systran system and from that date its translation service has developed and installed versions for a large number of language pairs for use within the Commission. Subsequently, the Commission decided to support the development of a system designed to be 'better' than Systran, which at that time was producing poor quality output, and began support for the Eurotra project – which, however, did not produce a system in the end... During the 1970s other systems began to be installed in large corporations. Then, in 1981 came the first translation software for the newly introduced personal computers, and gradually MT came into more widespread use. In the 1980s there was a revival of research, Japanese companies began the production of commercial systems, computerized translation aids became more familiar to professional translators. Then in 1990, relatively recently, the first translator workstations came to the market. Finally, in the last five years or so, MT has become an online service on the Internet.

1.3. Typology of systems and translation demands

We need to distinguish two basic types of system. Firstly there is the wholly automatic system that attempts to translate sentences and texts as wholes – that is to say, there is no intervention by any human user during the processes of translation. Because the outputs of these automatic systems are generally poor, companies and corporations need to provide human 'assistance' in order to improve quality. Second, as opposed to wholly automatic systems, there are various translation aids, which provide linguistic help for translators: most obviously, in the form of dictionaries and grammars but also (now most importantly) what are called translation memories, i.e. databases of previously translated texts to aid translators.

From the user's point of view the distinctions that are made by MT researchers between different types of automatic systems are largely irrelevant. For them, the MT system is a 'black box'. They are not concerned whether a MT system is transfer-

based or interlingua-based, whether it uses statistical data, whether it analyses sentences as phrase structures or dependency structures, and so forth. Users are involved with the workings of systems only – and sometimes not even then – for the compiling or augmenting of dictionaries used in the system, or for the processes of storing texts to be used in translation memory systems, or for preparing texts for MT systems.

For the user, what is more important is the way systems and tools are used. Any of the various types of MT system can be used for virtually any of the following basic functions.

(1) *Dissemination*: the production of translations of 'publishable' quality; not necessarily texts that are actually published but texts that are of that quality. Such texts are required usually by organizations and usually involve professional translators. The 'raw' untreated output from MT systems is inadequate, and publishable quality means human assistance; e.g. post-editing (revision) of the output text, pre-editing of the input, using a controlled language, or restricting the system to a specific subject domain. In general it has been found (through the experience of using MT) that the more the subject domain of an application can be restricted the more successful the system is, in terms of quality.

(2) *Assimilation*: the translation of texts for monitoring (or 'filtering') or skimming information, or the translation of texts for occasional users (e.g. non-specialist general public), where the 'raw' output from the system doesn't need to be edited; in other words, where recipients can accept poor quality as long as they can get an idea of what the text conveys. (In fact, we find that many MT systems are useful only in this function – their outputs are unsatisfactory as drafts for dissemination tasks.)

(3) *Interchange*: the communication between different languages by individuals, by correspondence, by email or by telephone. Here again the quality of the translation (and/or closeness to the original) is not so important, as long as people get the information they want, understand the message they receive, or manage to convey their intentions.

(4) *Database access*: the use of translation to assist in getting information from a database in a foreign language, one that the user does not well understand – i.e., these days this means mainly the use of translation aids for searching the Internet, for accessing web pages.

While dissemination and assimilation have been traditional uses of MT and translation tools from the beginning, their uses for interchange and databases access are recent applications which are growing rapidly, primarily with the growth of the Internet itself.

2. Fully automatic systems for dissemination

2.1. Using MT to produce good quality documents

The diagram in fig.1 illustrates the options available for the use of MT in a dissemination function.

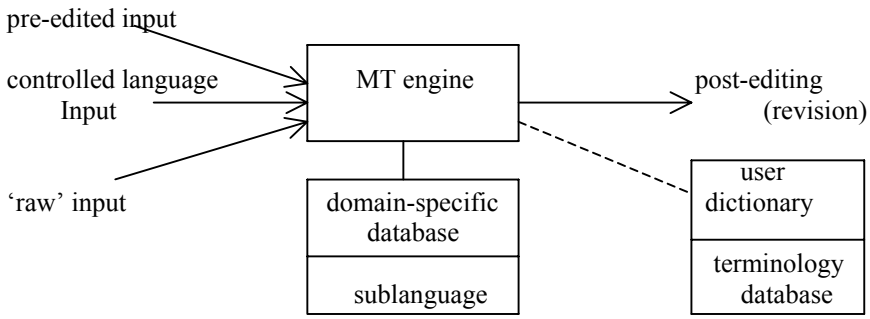

Fig.1: Human-aided machine translation

The MT system lies at the centre. It may be restricted (or adapted) to a particular domain or sublanguage – e.g. the language of medicine, which is different from the language of engineering, the language of art criticism, etc. – by means of the definitions and constraints specified in the databases of domain or sublanguage information. There is a MT engine (functioning as a 'black box'), into which text is put at one end and from which text is received at the other. The input can be either unedited ('raw') or it can be 'controlled' in some way, which means that either it can be 'pre-edited' by inserting various markers in the text to indicate how ambiguities or difficulties can be overcome, or it can be composed in a 'controlled language', a language designed to be regular and compatible in some way with a specific MT system. Whether the input is controlled or not, the output is invariably 'post-edited'; i.e. the text from the system is edited or revised by a human translator. Further human 'assistance' comes from the translator and/or colleagues through updating the dictionary of the system in order to improve the quality of the output, by including user-specific information and by augmenting lexical information from a database of 'approved' terminology. In corporation ('enterprise') and other commercial uses all of these possibilities (pre-editing, controlled language, post-editing, user dictionaries and terminological resources) may be found, or only one or two of them – but always there is some post-editing.

Large-scale translation using MT is cost effective – there are many large companies saving time and money with MT, but certain conditions must be present if it is to be possible. We are talking about technical documentation – not literature, sociology or legal texts – and, generally speaking, within a particular domain. Typical texts are internal reports, operational manuals, repetitive publicity and marketing documents. Operational manuals, in particular, often represent many thousands of pages to be translated, and are extremely boring for human translators, and they are often needed in many languages (English, French, German, Japanese, etc.). But companies want fairly good quality of output as well. Manuals are repetitive, there may be frequent updates; and from one edition to another there may be very few changes. Automation is the obvious answer.

To operate MT effectively, companies need adequate resources – not just the human beings involved (the translators, revisers, terminologists, computer engineers), but also a substantial body of translated texts as a basic corpus on which to design the system, and a good terminological database to ensure consistency of usage. They need to set down guidelines for an appropriate controlled language, and guidelines for post-editing. They need also to integrate the system with the technical writing and publishing sides of the workflow. The aim is to save costs, and perhaps staff as well – although the latter is a disputable and contentious matter.

What are the MT software requirements for corporations? Primarily 'enterprise systems' are client-server systems: a large server to house the MT system itself and PC-type clients linked on an intranet, often trans-national or trans-continental. The MT software has to be 'customisable' to the particular needs of the company. The system must come with a large basic dictionary – the larger the dictionary the better the output, generally speaking – plus substantial technical dictionaries, as close to the specific needs of the company as possible. User (or company) dictionaries will have to be made, and the system must have the facilities for easy dictionary creation and maintenance. Finally, the MT system must be able to be run on platforms compatible with those already used by companies (e.g. Unix, WindowsNT, Sun Solaris, Linux).

Systems covering the major Western European languages (English, French, German, Italian, Portuguese and Spanish) include Systran Enterprise, LogoMedia Enterprise, Reverso Intranet, SDL Enterprise Translator, and IBM WebSphere – all are client-server systems designed specifically for large company translation services. Another group of systems cover English and Far Eastern languages (Japanese, Korean, and Chinese), e.g. Fujitsu ATLAS, EWTranslate, IBM WebSphere, and Systran Enterprise. Other systems cover Arabic and English (AppTek's TranSphere) and Finnish and English (Kielikone's TranSmart). The dominance of English means that English appears in all systems, as source and/or target language, and other combinations are less common, e.g. French and German, or absent altogether, e.g. Dutch and Italian, Spanish and Japanese. Many languages are not represented at all in enterprise systems, simply because there is not a large commercial demand for those languages or for translation to and from them.

2.2. MT in the European Union

The European Union (previously European Communities) is one of the longest users of MT (apart from the US Air Force), and it is probably the largest user of MT. What are its systems used for? Some of the use is for the purposes described above – to produce translations which are post-edited by translators to publishable standard. In many cases, EU translators use MT output as the first draft of a translation (i.e. a pre-translation), which they then work on at a word processor. In most cases, MT is used for internal documentation. It seems to be rarely used by translators when translations are intended for external publication. However, translators are not now the main users of Systran translations (Senez 1995). They are Commission administrators, who receive many hundreds of documents which they need to know about; MT is used to get the rough gist of texts, and to decide whether it is worthwhile obtaining a human translation. Administrators also ask for the rapid post-editing of MT texts – by translators – which may then be destined for internal distribution within the Commission (Wagner 1985). A third use by administrators is as a draft for writing a document in a (non-native) language with which they are fairly familiar but in which they are not fluent. Finally, a recent development is the use of MT by Commission interpreters to produce translations of texts which they know are going to be reported on at sessions of the European Parliament or other meeting, and which they can refer to as a basis for their interpreting.

The Commission has now versions of Systran for many language pairs: English-French was first, then came French-English, English-Italian, followed by pairs involving German, Dutch, Spanish, and more recently Greek and Portuguese. The most common languages were covered first – for obvious reasons. In coming years, of course, the languages of Eastern Europe will be added – and work has

already begun on Czech and Polish. Demand for MT within the Commission has grown five times in less than 10 years (since the middle of the 1990s), and by over 20% per annum; and this rate may well increase in the next few years.

Quality has not improved so rapidly, but it has been found by the Commission that as long as translation can be restricted in subject matter or by document type to some extent, improvements in quality can be achieved. This can be demonstrated by the following example of translation of the same English text into French, firstly in 1987 and then in 1997 (Beaven 1998). It demonstrates the definite improvements that can be achieved by the EC 'adapting' the system to its own particular types of documents – and such texts are typical for the Commission.

> [English original] Since no request concerning changed circumstances with regard to injury to the Community industry was submitted, the review was limited to the question of dumping.
>
> [French 1987] Puisqu'aucune demande concernant les circonstances changées en ce qui concerne la blessure à l'industrie communautaire n'a été soumise, l'étude était limitée à la question de déverser.
>
> [French 1997] Comme aucune demande concernant un changement de circonstances en ce qui concerne le préjudice causé à l'industrie communautaire n'a été présentée, le réexamen était limité à l'aspect du dumping.

Another example, this time from English into Spanish, has a similar restriction of document type. The improvement in quality is readily discernible, although the result is still not 'perfect'.

> [English original] No formal list of supporting arguments was compiled but a number of points were common to the papers and discussions, including the following: ...
>
> [Spanish 1987] Ninguna lista formal de mantener las discusiónes fue compilada pero varios puntos eran comunes a los papeles y a las discusiones, con inclusión del siguiente: ...
>
> [Spanish 1997] No se compiló ninguna lista formal de argumentos favorables sino que varios puntos eran comunes a los documentos y a las discusiones, incluida la siguiente: ...

2.3. Post-editing MT output

For texts which are intended for dissemination, post-editing is always considered to be essential (Allen 2003, Kring 2001). But what are the circumstances that give rise to 'errors' which need to be post-edited? The main errors arise, of course, from the difficulties computers have with many aspects of language (ellipsis, pronouns, coordination, to mention just a few) and, in particular, with handling complex sentences – long sentences of more than one clause tend always to be translated less successfully than short single-clause sentences.

In some cases, the reasons for mistranslation may be relatively trivial, but post-editing is still required. If the original text includes a misspelling – which, despite spell checkers and grammar checkers, still happens – the word will not be recognised and this may well affect translation of the rest of the sentence. If the original contains a typographic mistake (e.g. the use of from instead of *form*), there is going to be a problem with translation. Missing punctuation can also cause problems. For example *The Commission vice president* might be incorrectly translated as *Le président du vice de la Commission* (the original should have a hyphen between *vice* and *president* in order to produce *le vice-président de la Commission*.)

Some problems of interest to a Spanish audience are the following. Prepositions are always problematic for MT systems. From this Spanish example:

> …el desarrollo de programs de educación nutricional...

the MT system might produce:

> …the development of programs of nutritional education

which a post-editor would correct as:

…programs in nutritional education…

i.e. English prefers *in* after this type of noun.

Prepositional verb phrases are equally problematic. Spanish *para* would normally be rendered as *in order to*, but in the following example:

…el procedimento para registrar los hogares…

and MT output:

... the procedure in order to register the households

English would prefer a prepositional gerundive:

…the procedure for registering households

It is almost a matter of English style, and notoriously difficult to get right in a MT system.

Inversions are another problem. From:

…la inversión de la Argentina en las investigaciones de malaria

a MT system might produce:

…the investment of Argentina in the research of malaria

Although this is perfectly comprehensible, the post-editor would probably change it to:

... Argentina's investment in malaria research

Then there is the problem of the common Spanish constructions with reflexive verbs. From this sentence:

Se estudiarán todos los pacientes diagnostocados como…

the MT system might well give:

There will be studied all the patients diagnosed as…

English prefers a passive construction:

Studies will be done on all patients diagnosed as…

Another example of a reflexive:

En 1972 se formuló el Plan Decenal de Salud para las Américas.

The MT system might produce:

In 1972 there was formulated the Ten-Year Health Plan for the Americas

Which the post-editor would probably change to:

The year 1972 saw the formulation of the Ten-Year Health Plan for the Americas.

Such are the kinds of post-editing operations required if the text is to be of publishable quality. In some cases, it would be possible to correct theses regular errors semi-automatically by using macros in the word processor system, and this approach (first proposed for post-editing of output at the Pan American Health Organization – Vasconcellos 1986) is adopted increasingly in large translation services. If the document is not to be published then the uncorrected versions might well be acceptable because readers can still understand the meaning – as indeed in the cases illustrated above.

The amount of post-editing can vary considerably (Allen 2003). The more texts are 'standardized', the more they are full of jargon and clichés, the more the text is mundane and uncreative, the more accurate will be the MT output and the less correction by post-editing is required. Machine translation works best on standardized input. Creativity is penalised. Unfamiliar word combinations and sentence constructions result in poor MT versions. The more uncreative a text, the better the results.

Usually post-editing is done by translators. This was not initially foreseen by MT pioneers – they thought that anybody could do revision if they knew the target language. But it was quickly realised that revisers do need to know the source language in order to do revision. Post-editing depends largely on bilingual skills acquired over time, and translators have these skills more than non-translators.

However, translators learning to do post-editing have to persevere: initially post-editing takes longer than translation from scratch. The advantages of translators as revisers – as opposed to non-translators – is that they can maintain quality control. Consistency of terminology can be maintained by the MT system, repetitive and familiar matter can be left to the MT system, but the linguistic quality needs to be maintained by the human translator. The disadvantages of using translators as post-editors are that translators are constantly correcting the same trivial mistakes – MT systems are consistent: they make the same mistake every time – and the style of MT output is usually far too oriented towards that of the source language – the translator is constantly changing sentence structures. It is irritating to be doing this all the time. Perhaps the answer might be specifically trained post-editors, people with some knowledge of translation but who are prepared to learn the skills needed to specialize as good post-editors.

2.4. Preparing input for MT

While post-editing can correct MT mistakes, many can be avoided by human preparation of the input texts. How can the input be adapted to the MT system? The first idea was that the original documents could be written by authors in a language designed for a MT system, in what was called 'MT-ese'. It was rather an unrealistic idea, but nevertheless seriously proposed. More acceptable was pre-editing by the operators and users of the system. Pre-editing can be done at various levels:

(1) to indicate whether a particular word has a particular function (e.g. *convict* as noun or verb). Reducing functional ambiguity greatly simplifies MT analysis of the sentence.

(2) to indicate whether a particular name is to be translated or not. It could be done to prevent, for example, *John White* coming out as *Johann Weiss*.

(3) to indicate the boundaries of compound nouns. For example, we want *light bulb* to be translated in French always as *ampoule* and not as *bulbe léger* or *oignon léger*. To ensure this the compound can be enclosed in brackets.

(4) to insert punctuation. For example in a sentence such as *There are he says two options…* we may need to ensure that *he says* is treated as an embedded phrase. It could be enclosed in commas or bracketed, e.g. *There are, he says, two options…* or *There are (he says) two options...*

(5) to split long sentences into shorter ones. This is more drastic pre-editing, done because experience shows that MT cannot deal well with long complex sentences (and this advice is often given to purchasers of personal MT systems). It is in effect a form of 'controlled language' (next section).

In theory, pre-editing does not require knowledge of the target language, but in practice if it is useful to know, for example, that in a specific target language a particular phrase is likely to be misinterpreted (as in the *light bulb* example) or, on the other hand, to know that there is *not* likely to be misinterpretation and that no marking is necessary when translating into the intended target language.

2.5. Controlled language and MT

A system can be adjusted to a specific subject domain in two ways: either by designing the system itself specifically for a particular sublanguage (Somers 2003d) – as the Météo system was – or, much more commonly, by restricting the range of vocabulary and the grammatical structures of texts input to the system. It is the approach favoured by corporations needing to translate into a large number of languages. Controlled languages (CLAW 1996, EAMT-CLAW 2003, Nyberg et al.

2003) enable input texts to be standardized and consistent, ideally unambiguous from the viewpoint of a MT system, in order to minimize the problems of lexical selection and structural change, and therefore to produce better quality output which may need little or no post-editing.

Examples of the kinds of rules to be followed by writers in controlled languages (or for editors rewriting texts into a controlled form) are:

(1) Always use acceptable ('authorized') terminology. In the case of a car manufacturer, there might be a rule specifying the use of *windscreen* rather than *windshield*

(2) Use only approved senses of ambiguous words, e.g. *follow* should be used only in the sense of *come after* and not of *obey*. (To 'follow a command' means to obey a command)

(3) Avoid ambiguous words like *replace*, since it can mean either 'remove a component and put it back again', or 'take the component away and put in a new one'. Similarly, the word *appear* can mean *come into view*, or *be possible*, or *show*. The rule may state that *appear* should be avoided altogether.

(4) Use one topic per sentence; e.g. one command or instruction at a time. (This reduces the complexity of sentences.)

(5) Do not omit articles – otherwise there is ambiguity

(6) Avoid phrasal verbs such as *pour out*, *put in*, *look up*, *look into*, .. etc. (The MT analysis of such verb forms often causes problems.)

(7) Do not omit any implied nouns (e.g. subjects omitted in coordinated structures).

(8) Use short sentences – often guidelines for controlled languages specify a maximum of 20 words per sentence.

(9) Avoid coordination, since it always causes problems in MT systems, e.g. in the phrase *homeless women and children*, the adjective probably modifies both nouns, but in *pregnant women and children* the adjective modifies only the first noun.

As an illustration of the effects of controlling source language input, take the sentence:

After agitation, allow the solution to stand for one hour

The phrase *after agitation* does not indicate the object of the action. It is better to be more precise and to avoid the ambiguous *agitate* by rephrasing as:

If you shake the solution, do not use it for one hour.

Another example illustrates the rule about keeping to one topic (or command) per sentence. Instead of:

It is very important that you keep all of the engine parts clean and free of corrosion.

The controlled language form could be:

Keep all of the engine parts clean. Do not let corrosion occur.

The idea of controlling the language of MT input is an old idea. Stuart Dodd proposed 'Model English' in 1952 for the writers of texts submitted to MT. At this time, restricted languages were already in use for air traffic control. In the last twenty years there have been many controlled languages proposed and implemented – many not specifically for MT, but because of the recognised advantages of unambiguous language in many situations. Some examples are: AECMA Simplified English for car manufacturing, CFE (Caterpillar Fundamental English) and later CTE (Caterpillar Technical English) for use with the MT system developed at Carnegie-Mellon University for the Caterpillar Corporation, MCE (Multinational Customized English) used by Xerox Corporation in its application of the Systran system, and PACE (Perkins Approved Clear English, developed for an engineering firm using the now

defunct MT system Weidner. There are a number of companies which develop customized MT systems with controlled language input, e.g. Smart Communications (New York) for many multinational organizations (Citicorp, Chase, Ford, General Electric, Canadian Ministry of Employment, etc.), Xplanation (Belgium) using the METAL system, VTT Information Technology (Finland) using the WebTranSmart system, and ESTeam (Greece) – all covering a wide range of European languages.

2.6. In-house MT systems

A number of organizations have developed MT systems for their own use. In so far as the documentation is specific to relatively narrow subject domains, these systems have similarities with controlled language applications: the input is not necessarily restricted, but in effect the range is limited – thus MT is made somewhat easier. Examples are found in many countries. Some Japanese examples are: the system at JICST (Japan Information Center for Science and Technology) dedicated to the translation of abstracts of science articles from English into Japanese; the NHK system for translating news broadcasts into English; and the systems at IBM Japan and CSK for translating their own documentation between Japanese and English. In Europe we have the PaTrans system in Denmark for translation of patents from English into Danish (based on research from the Eurotra project), a one-man development at Hook and Hatton in England for translating chemistry texts from Dutch into English, the Linguanet system from ProLingua for translating police and air traffic control messages between various European languages – police language is often limited in subject range, and air traffic control language has to be restricted as users do not necessarily know English very well. In North America there are a number of restricted systems. One has been developed specifically for TV captions (from English into Spanish), involving transcription of the spoken language and segmentation of text (with problems of word recognition) and needing to be very robust – it is nearly real-time system, since although subtitles of TV programs are usually prepared in advance, they are often done close to the time of broadcasting. Another is a system for military MT in restricted domains, the DIPLOMAT system from Carnegie Mellon University, developed for Croatian, Spanish, Haitian Creole and Korean (and probably currently for Arabic in the Iraqi situation). It is a very restricted system, not much more than a phrase dictionary. More recently, on the same lines, there is Phraselator, which as the name implies, translates standard phrases from and to English and a large number of languages: Albanian, Arabic, Bengali, Cambodian, Chinese, etc. – mainly for military uses, but also to aid tourists.

2.7. Lexical resources

Operational systems (whether purchased from commercial MT companies or developed in-house) must all have substantial dictionaries that have to be kept continuously up to date. Until recently, creating and updating dictionaries was an expensive and labour-intensive operation. Every entry had to be entered separately. In addition, the information required for each dictionary entry went well beyond the information readily obtained from printed dictionaries. MT systems required details about grammatical categories (usually more refined than the traditional noun, adjective, adverb, etc.), morphological variants (e.g. all possible noun and verb endings), the syntactic contexts of word usage, information about semantic constraints on the word occurrences, and of course all the possible translation alternatives and the circumstances in which each might be selected. The reason was that computers were then designed to be operated by following precise lexical and grammatical rules, and

existing dictionaries were unable to supply the detail required. Now, however, more recent MT systems operate on shallower syntactic information, and rarely incorporate complex semantic rules for disambiguation – instead there is much more use of information about word cooccurrences derived from text corpora. The depth of detail is much less than in the past, and it is now possible to extract information for compiling dictionary entries directly and automatically or semi-automatically from bilingual corpora.

For the company purchasing MT systems there is yet another factor to be considered. Are the basic dictionaries to be supplied wholly by the vendor, or should they be compiled by the customer? Since most companies have their own peculiarities of terminological usage, the answer is invariably that the corporation itself must be heavily involved in creating dictionaries – which means a substantial financial commitment.

It is therefore an increasingly important issue for the practical utilization of MT how and whether lexical resources can be acquired more easily and cheaply. Although many systems are effectively restricted to use within specific subject domains, the definition and the size of those domains are not easily determined. In any case, texts stray over neat subject boundaries; and many texts of interest to a company lie outside the strict domain of its business. In practice, a company will need dictionaries covering many different subject areas (close to and distant from its central concerns). As well as using familiar printed resources (or their electronic equivalents on the Internet), the developers (or adapters) of a system turn also to electronic text corpora from which they may be able to extract relevant lexical information. It is the availability of large text corpora (whether bilingual or not) on the Internet that has encouraged in recent years the investigation of methods for the automatic extraction of terminology from textual databases of all kinds (cf. LREC 2000) – and a number of non-governmental organizations have been established to collect and provide access to such resources (e.g. ELRA and LDC).

Once the terms have been acquired for a company system it is not, however, the end of the story: the terms have to be validated and authorized – the quality of MT output depends to a great extent on the quality of the system dictionary, and furthermore, companies have often their own terminological preferences (as seen in the control of language, above). Finally, since vocabulary in all areas is constantly changing, dictionaries have to be regularly updated – there has to be a regular trawling of lexical and terminological resources and a continuous programme of creating, maintaining and validating MT lexical information. There are now a large number of commercial software packages which perform terminology extraction and support terminological management.

3. Translation aids for dissemination

3.1. Tools for translators

At this point we turn to the computer aids for translation – also referred to as Machine aided human translation (MAHT). Rather than investing in a MT system (with all the concomitant commitments and expenses outlined above), a company may well prefer to provide translators with a range of computer-based tools.

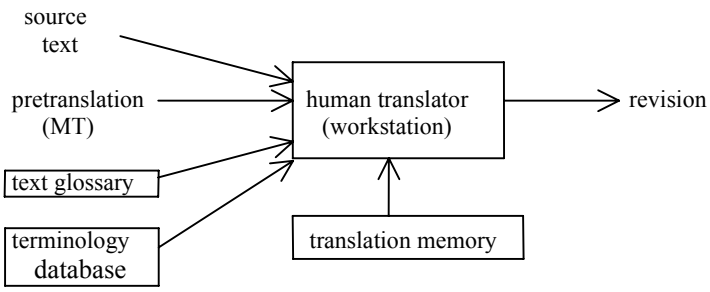

Fig.2: Machine-aided human translation

In this scenario at the centre is not the MT engine (as in fig.1) but the human translator at a workstation (computer). Available to him or her are various optional computer facilities, comprising typically a terminology database, a facility to create a text glossary (a list of words in the source text and their translations), access to 'provisional' pretranslation from a MT system, a translation memory, and facilities to revise texts after translation. These days, all these options and resources are integrated as a translator's workstation (Somers 2003b), enabling the translator to select the aids he wants to use and with which he or she does the actual translation. Unlike the MT set-up (fig.1 above), it is not the machine which is at the centre. Here the translator is not subservient to the machine, but fully in control.

The translator's workstation is a relatively recent development, which has come about with changes in the way translators work (Melby 1992, Hutchins 1998). The development of electronic termbanks, the increasing need to adhere to terminology standards, the overwhelming volumes of translation, and above all the development of facilities for using previous examples of translations has meant that translators could see the practical advantages of computerization in their work. In addition, translators were increasingly using personal computers; they were familiar with the benefits of word processing and desktop publishing; their clients were providing texts for translation in electronic form and they wanted the results transmitted electronically, quickly and fully formatted.

The answer has been the translator workstation, made possible by two important developments. The acquisition and management of terminology has become essential for any translator involved in translating within a particular subject domain, whether as an individual or on behalf of a company. The translator needs to extract terms, definitions, and examples of context from either existing terminology databases or from bilingual text resources. The information needs to cover grammatical and semantic detail, provide authorized definitions, specify any relevant standards, give translation equivalents, and supply examples of usage. The information is stored in a terminology database (accessed word by word as required, or collectively as a text glossary – a list of words and example translations for one specific text). The terminology information has to be kept up to date, corrections have to be made, etc. – organization of the data has progressed from the old card index systems to computer programs, such as MTX (from Linguatech, the first) and MultiTerm (from Trados). Not only is it more efficient for the individual translator, but computerization has enabled translators and organizations to share terminological information – now using agreed exchange standards (MATER, etc.).

The second important development – often, in fact seen as the defining development – has been the appearance of systems for storing large corpora of bilingual texts, and for extracting from them examples of translations corresponding to phrases and sentences in the source text which the translator may be working upon.

These databases, now known universally as 'translation memories' (Somers 2003c), were in fact a by-product of research on statistical methods of MT (Hutchins 1998). In essence, translators can now easily see how a word, phrase or sentence had been translated in the past (perhaps by the translator himself/herself, or by another translator) – there is no need to do laborious searches through printed documents. In particular the translator can find how individual words and phrases have been translated in different ways in different contexts. Furthermore, in the case of terms which have to be translated consistently in the same way, the translator can now easily find out what the approved version should be. Above all, however, the availability of previous translations in electronic form means that large portions of texts which have already been translated – in whole or in part – need not be re-translated (at great cost). From the translation memory (TM) the translator can extract those sentences of a source text which remain unchanged and re-produce the earlier translations. In circumstances where texts are frequently updated, sometimes with little change (as in many operational manuals for machinery), the translation memory facility can reduce costs substantially. However, translation memories are not without their own problems. Translators find often that inexact examples are as important for them as exact repetitions, but most TM systems have difficulties with 'fuzzy matching' – either too many irrelevant examples are extracted, or too many potentially useful examples are missed. Other difficulties encountered are the presence of redundant examples – unnecessary either because there are already many similar examples in the database (and the translator is thus presented with too many very similar sentences), or because the sentences contain rare or untypical translations which serve only to confuse.

Although the translator workstation integrates a variety of computer-based translation tools, such is the central role of the translation memory that they are now often referred to simply as 'translation memory systems'. In general, translator workstations are most suitable for translators in large corporations and large translation agencies, where there are large volumes of texts focussing on a particular subject field or particular range of products within a specific field, where much of the documentation is repetitive, and where consistency of terminology is most important. Apart from the integration of TMs and translation tools, many workstations now include MT programs as an further option, so that translators may be aided by a rough MT version if, e.g., fuzzy matching in a translation memory has failed to give results. Translator workstations (or TM systems) can also be successfully linked closely to (computer-based) authoring and publication systems. It is now common for major companies to seek further reductions in the production costs of their documentation by integrating technical writing and translation processes.

Today there are many commercial translator workstations available. The oldest and best established are the systems from Trados, STAR (Transit), and Atril (Déjà Vu). Others include the systems from SDL (SDLX), Multilizer, Terminotix (LogiTerm), Champollion (WordFast), MultiCorpora (MultiTrans), MetaTexis, and so forth. Facilities in them are very similar, and the competition for the increasing demand is intense.

3.2. Localization

Probably the largest users of computer aids for translation are found in the field of software and web localization. This is now a major sector of the translation industry, with its own organization, LISA (Localization Industry Standards Association), holding a large number of conferences, six or seven every year. A recent

report from ABI estimated a market of over 6 million dollars by 2006 (roughly half for software localization and half for webpage localization). By localization is meant the adaptation of products for a particular national or regional market, and the creation of documentation for those products (Esselink 2000, Esselink 2003, Sprung 2000). Localization means not just translation into the vernacular language, it means also adaptation to local currencies, measurements, and power supplies, and it means more subtle cultural and social adaptation. The incentive for computerization is the demand for the localization of publicity, promotional material, manuals for installation, maintenance and repair, etc. These must be available in the country (or countries) concerned as soon as the product itself is ready – often in a matter of days, not weeks. Those involved in localization are under huge pressures to produce localized documentation at short notice. Since the changes from one version of a product to another are often quite small there may also be few changes in the documentation. Indeed there may also be relatively few between the documentation of one product and the documentation of another from the same company. With such high levels of repetition the translation memory becomes obviously a vital resource, and so TM is in heavy and increasing use in the localization industry.

In addition to translation memories and other translation aids there are now a range of other ancillary computer-based systems to support localization, e.g. tools for project management, document control, quality assurance.

A recent development is the appearance of software for translating webpages. Companies must now maintain high-profile presence on the Internet, in order that they remain competitive. For multi-national companies this also means that information on their websites must be made available in multiple languages. One solution is to refer users to online MT services (see 4.1 below) – but for many reasons this is unsatisfactory. Another is to engage a localization agency to translate every webpage. A third option which is increasingly adopted is to integrate one of the automatic webpage localization systems offered by many of the vendors of MT systems. Examples are ArabSite, IBM WebSphere, InterTran Website Translation Server, SDL Webflow, SystranLinks, Worldlingo.

3.3. Convergence of MT and TM systems

Various similarities will have been noticed in the descriptions above between the use of MT systems in large corporations and the use of translation tools. In recent years, we are beginning to see the convergence of these various facilities; increasingly we find systems that straddle the categories. On the one hand, workstations often include MT components, and on the other, MT systems are starting to include TMs as system components. Some localization companies use both MT and TM – each having their advantages and disadvantages. The motivation for convergence is that very many facilities are exactly the same in both: terminology management is needed in MT as much as it is in TM systems, both MT and TM systems need to be closely integrated with authoring and publishing, both demand quality control, methods for lexical acquisition, and facilities for webpage translation. Above all there is the common aim: companies use MT and TM to produce quality translations (publishable quality documents), and use the skills of translators to achieve them. As yet, however, we still find the two (MT systems and translation workstations) at work in parallel; there are still uncertainties about how they might be fully integrated. One requirement must be for MT systems and/or TM systems to be capable of dealing with text segments smaller than sentences (e.g. noun phrases, verb phrases, adjectival and prepositional constructions, etc.), finding their equivalents in the target language and

combining them in correct (idiomatic) target language sentences. For this, it may well be that the appropriate computer facilities for both machine and human use will emerge from current research on what is called example-based machine translation (EBMT) – see Carl and Way 2003.

3.4. Office systems

Before leaving translation for dissemination, mention should be made of the systems available for the small-scale use of MT, i.e. the use by individual (independent) professional translators. They need access both to translation databases (TM) and to some kind of terminology management, because they are often producing translations regularly for individual customers. They need integration with their other IT equipment. They want to save costs, and they need easy post-editing facilities. In many respects, translator workstations would seem to meet these requirements, but in general they are still too expensive for most independent translators. So these translators want the functions available to large corporations, but they want them on their own personal computers (PCs) on less expensive set-ups. There are in fact a large number of vendors who produce systems for this market, either by downsizing their client-server systems or by upgrading (improving) their cheaper PC systems (i.e. those for 'home use', to which we will come to shortly). Vendors refer to these systems as 'office' or 'professional' systems. These systems are specifically designed for professional translators since they incorporate features not present in 'home' systems, and they do not include the full range of facilities of corporation ('enterprise') systems. As a result, these systems could be attractive not just to individual translators by also to others, such as small companies that cannot afford translator workstations or enterprise systems, or smaller translation agencies and even by some 'home' or personal MT users.

In this category, there is a large range. Many MT vendors cover English to and from French, German, Italian and Spanish – such as Systran, LogoMedia, Reverso, Transcend (SDL). Others specialize in one or two language pairs. For Spanish there are the systems from the Pan American Health Organization (downsized from mainframe in-house systems), i.e. SPANAM for Spanish into English and ENGSPAN for English into Spanish, and the ESI Professional from WordMagic (English-Spanish in both directions). For English/Italian there is Synthema's PeTra; for English/German Linguatec's Personal Translator PT, and Langenscheidt's T1; and for Russian/English there are systems from ProjectMT (@promt) and Lingvistica (PARS). For Japanese there are systems from Hitachi (HICATS), Toshiba (Honyaku Office), Oki (Pensee), LEC (LogoVista), Cross language (Transer), and Systran. CITAC, LogoMedia and Systran have Chinese to English systems, and English to/from Korean is covered by LogoMedia, LogoVista and Systran systems. In addition, TranSphere (from AppTek) sells systems for translating from English into Arabic, Chinese, French, Japanese, Korean, Persian and Turkish. Systems between French and German, French and Spanish, German and Russian, and other pairs not involving English are now not as rare as they were once, e.g. Reverso, Systran, T1, PARS. Finally there are the Hypertrans systems designed specifically for the translation of patents between a wide range of language pairs. For details of these and other systems see the *Compendium of translation software* (Hutchins et al. 2004).

4. MT for assimilation, and Personal MT systems

4.1. MT for information purposes

So far we have looked only at MT and translation aids for the production of publication-quality translation, systems which are used primarily by translators in large organizations or as independent professional translators. The other main use of MT is for assimilation, for getting the gist (essence) of the basic message of a text. The recipient does not necessarily require good quality. The main requirement is immediate use. However, the output must be readable, it must be reasonably intelligible for someone knowing the subject background. In many cases, the 'raw' MT output may be adequate, but in others some 'light' or 'rapid' post-editing may be done (e.g. Wagner 1985).

Major users of unedited MT have been large multi-national organizations, e.g. the European Commission, and in particular the intelligence services, e.g. the CIA, the National Air Information Center, and other government bodies in the USA and elsewhere; and indeed, systems operating for this purpose have been in use since the earliest days of MT – the CIA was the main sponsor of research at the Georgetown University in the 1950s and 1960s, the IBM Translator was used by the US Air Force, and so too was the Systran system which replaced it (Henisz-Dostert 1979). It was also the primary purpose of MT research in the former Soviet Union. Today there is intense interest in translation from Arabic (both spoken and written), ideally in conjunction with information extraction and summarization systems (see section 5 below.)

Today, however, there is another even larger market for assimilation systems. This is the translation needs of members of the general public, often with little or no knowledge of languages, who want translations on their own personal computers – most probably of texts obtained from the Internet, or of webpages found when searching Internet resources.

There is a huge number of PC systems for this kind of 'home use', covering many – but by no means all – languages. There is still a major emphasis on European languages (English, French, German, Italian, Portuguese, Russian, Spanish); and then on English-Japanese, English-Chinese, English-Korean, and English-Arabic – for examples see the *Compendium of translation software* (Hutchins et al. 2004). Once we go beyond these languages, however, there are few systems available, e.g. for languages of Africa, the Indian subcontinent and South East Asia. It is also noticeable that nearly all are for languages to and from English – an unfortunate consequence of the hegemony of the English language and the American economy. Many of the systems have special facilities for translating web pages (a few are specifically for this purpose). As for quality, there are quite a few which produce abysmal results, but there are also some systems (mainly those downsized from older mainframe systems) which do produce quite acceptable unedited translations – good enough for assimilation, but not otherwise.

The free MT services on the Internet are now familiar to almost everyone. Most believe they appeared only in the last three or five years – in fact, the first one dates from the mid 1980s: the Minitel service using the Systran system on a network in France; and in 1994, Compuserve started providing free translations in 1994. The best known is now Babelfish from AltaVista. As with personal MT systems we find that services concentrate on western European language to and from English: French, German, Italian, Portuguese and Spanish. A few offer French to/from German, and French into Spanish. Then there are systems for English, Russian, Polish, and

Ukrainian (e.g. PARS, PROMT, Poltran), and of course English into and from Chinese, Japanese, and Korean. For other languages, Arabic is quite well provided; but, as before, other languages are neglected. The picture is constantly in flux; services come and go within months; some produce relatively good output (e.g. those based on Systran, Transcend, Reverso and Promt), but many are little more than dictionaries giving therefore virtually word for word 'translations' with hardly any lexical selection mechanisms. There are in addition to the free online services some online MT services where a fee can be paid in order to have the output post-edited. The charges are obviously dependent on the availability and the demand for human translators. In general, these services are designed for companies which do not have their own translation facilities, or which may be considering the acquisition of MT or translation aids. Again, the language coverage concentrates on Europe and the Far East.

A recent trend is the appearance of hand-held (PDA) systems, mainly from the Ectaco company. A large range of languages is covered, but the sizes of the dictionaries are small, and so the quality may be doubtful. Some providers of these devices admit that they are just dictionaries, but they may be adequate for tourism. Lastly, there are some vendors (e.g. Linguatec) who have adapted their personal MT systems to use with SMS and mobile telephone, either stand-alone or in most cases operating through contact with remote proprietary databases.

The wide availability of free translation of webpages makes it possible for companies and organizations to reach potential clients and customers who are unfamiliar with the language of their websites; and many organizations provide links to such services for users to obtain translations of their websites. While this may show awareness that language barriers do need to be overcome – e.g. that not all Internet users know English well – it would seem that few of them are aware of the poor and unsatisfactory quality of website translations, the often poor usability of online services for non-English speakers, and many factors contributing failures of communication and frustration (Gaspari 2004). They are clearly unaware of the potential damage to the image of their own company and organization; evidently, many believe that the availability of online translation absolves them from any responsibility towards foreign-language users of their websites – in particular that they need not invest in web localization (whether through localization agencies or through web localization software section 3.2 above). It is in fact another instance of a failure to distinguish between translation for dissemination and translation for assimilation. Organizations with websites ought to aim for good quality dissemination of information, and not be content with low-quality semi-comprehensible 'translations' over which they have no control.

4.2. MT in the marketplace

There are features of the commercial MT scene which indicate some fragility in the current situation. Like other computer software, but perhaps to a greater extent, very few systems are sold from stores; they have to be purchased directly from the manufacturers – supplied on CD-Roms or from downloaded files. Clearly, MT is not a mass market product; it is still an unfamiliar product for the general public, and purchasers are venturing into unknown territory. They are largely ignorant about what they may expect from MT and translation aids. To add to purchasers' problems, vendors use confusing terminology. On the one hand there are vendors who label as 'translation systems' products which consist of little more than electronic dictionaries, i.e. make no attempt to translate sentences. And on the other hand, there are vendors

who label (modestly or honestly) as ‘computer-aided translation’ systems software which can produce good quality translations without human intervention.

For potential corporate purchasers, there is in fact a confusing use of the term ‘computer-aided translation’ itself. Sometimes it refers to human-aided MT, sometimes to machine aids for human translation. It is true that vendors may occasionally mean both, but usually they do not. Finally, there are confusing terms for systems that combine or include MT and various translation tools. As mentioned earlier, translation memories (TM) can be either independent products, or they can be components of a translator workstation, or they can be part of a MT system.

More problematic for MT is the question of what users expect from systems (particularly from personal MT systems). Generally they expect high quality (equivalent to that of human translators), but what they usually get is low quality. Most of the criticism of MT products, whether in specialist or in general magazines, emphasizes the inadequacies of their translations – criticisms often made by people knowing something of two languages but not as translators. Purchasers of MT systems need to be made aware that MT should be used only for certain types of texts within particular circumstances; they should not expect MT systems to translate everything. It is true that responsible MT vendors do give advice to users, such as breaking down long complex sentences into shorter ones, avoiding parentheses and ellipsis, etc., but ignorance of source and/or target languages by many users makes such advice difficult to implement, and the results continue to be poor.

Today users come to expect computer software to improve significantly over time – the next version of a piece of software is expected to be better than the last one. The trouble with MT is that this is rarely the case; in particular there may be virtually no improvement in the quality of output from one version to the next. It may well be true to say that over the last ten years there has been hardly any improvements in MT quality for many language pairs – see below (section VI). (In fact there is not a great deal of incentive for the commercial developers to improve their systems because they have enough sales without.) A further expectation of users must be that there more languages will become available – again the problem in the commercial field is that there is undoubtedly more money to be made from the major languages of the world than other languages in, say, India (e.g. Bengali, Hindustani, Gujarati), Africa (e.g. Swahili, Fulani, Yoruba) or South East Asia (e.g. Indonesian, Khmer, Burmese).

It is still a market where there are many potential purchasers who are quite unfamiliar with translation and even less with the range and variety of MT products available. What is urgently needed are consumer reports of MT systems, so that potential purchasers of systems can know what they are buying from a particular package, what its strengths and weaknesses are, what it should not be used for and what it might be used for. The MT community as a whole needs to set up some standards, benchmarks, consumer reports, etc. They do not exist at present. It must be admitted that the evaluation of MT is extremely difficult (White 2003), and to do this in a straightforward way, easily understood by the general public, is even more difficult, but it ought to be attempted.

The fragility of the MT market is illustrated by the disappearance of numerous systems in the recent years. Some once very popular systems are no longer available. The Weidner system marketed in its Japanese version by Bravis was once selling in thousands in Japan, but the owner decided it was not making enough money and he closed it down. Intergraph was another once very popular system was bought up by another company and then withdrawn; it is now no longer available. The Transcend system originally sold by Transparent Language was bought by SDL, but it is now

marketed at a very low profile, so it is little known in the marketplace. Globalink and Microtac were two companies which merged and were then absorbed by Lernout & Hauspie, which (as is well known) subsequently went into liquidation. Logos Corporation sold a very popular system for corporations; but it too ran into financial difficulties. A system for Scandinavian languages (Winger) seemed to have good prospects, but it folded as well. The well-regarded Japanese system Pensee from Oki has also been a victim of the market. The most recent departure has been Sail Labs which produced an impressive German-English-German 'enterprise' system (Comprendium) but it too has run into financial problems.

The basic problems for vendors are the low profit margins and the slow quality improvement. There are in fact very few obvious differences between rival systems – in terms of facilities, cost overheads, and translation quality, and even language pairs. The whole situation is further confused by the very existence of free online systems – people will ask themselves before buying a system why they should be paying when they can get translations free. They are unaware that a PC system may in fact produce better results than online systems – although unfortunately even this is not always true!

5. MT for interchange and database applications

Closely related to the use of MT for translating texts for assimilation purposes is their use for aiding bilingual (or cross-language) communication and for searching, accessing and understanding foreign language information from databases and webpages. Any of the systems already mentioned can be used for these functions; and indeed both 'enterprise' (client-server) and 'professional' systems are used in this way. Some have incorporated special facilities for searching and translating foreign language webpages, and most provide for the easy translation of email messages.

It is probably true to say that one of the main applications of personal MT ('home') systems is the translation of correspondence (including electronic, i.e. personal, emails) and the translation of web pages. Most of the PC systems mentioned above have specific facilities for email and webpage translation – the translations are in principle no better or no worse than for other texts, but their use for these tasks has been made easier. Email translation is actually quite likely to be worse than for well-written documents, since the systems themselves were not designed to cope with the ungrammatical and idiomatic usage found in much email communication.

The use of personal MT systems for translating into an unknown (or poorly known) foreign language may not be advisable – most reputable vendors caution users against doing so – but it is known that many purchasers of PC 'home' systems use them to translate letters, emails and other short documents into unknown languages – with consequences that can be imagined, but are rarely reported or admitted.

Apart from emails, there are many areas of interchange which involve translation of some kind. There is much business correspondence that needs to be translated. As yet MT systems have not been designed for this task, although in the future we may expect to see special-purpose software able to translate the typical style of business letters, often almost formulaic in nature. Likewise we may expect software designed to assist writers to compose letters in a foreign language. There has been quite a lot research on this application of MT, but there are no commercial systems as yet. Another example of interchange is translation into sign languages for deaf and hearing impaired people – this is just research at present, but it seems likely that systems for this function will be operating fairly soon.

Above all, there is oral communication involving translation. The translation of spoken language has not been mentioned so far; it is an area of intensive research (e.g. Kurematsu and Morimoto 1996, Wahlster 2000, Woszczyna et al. 1998), but there are still no MT systems on the market which deal with speech. Such are the problems (incomplete sentences, ellipsis, hesitations, backtracking, interruptions, implied contexts, etc.) that any systems for speech translation will inevitably have to be restricted to narrow domains, e.g. telephone bookings of hotel accommodation. In a few years perhaps there may well be some such highly restricted speech MT systems available – they will probably not be top quality, but they should be usable. Whether speech translation will later become more wide spread, with ever-looser restrictions on contexts, domains and speakers, is very much an open question. What is certain is that it will not come in the near future. In parenthesis, it may be remarked that cross-language interchange in the form of (near) simultaneous interpretation is most unlikely at any time, although (as mentioned above) interpreters can, and obviously do, make use of rough drafts from MT systems to aid them in conference and meeting interpretation.

Although we do not yet have speech translation, we do have systems with voice input and output, i.e. where users speak into the system, the spoken word or sentence is converted into text, the written text is translated into another text, and the system then produces spoken output. Quite a few PC systems provide voice input and output – usually voice output only. Examples are: Al-Wafi (English output), CITAC (English output), ESI (Spanish/English), Korya Eiwa (Japanese/English), Personal Translator PT (German/English, option in Office version), Reverso Voice (French/English, French/German), TranSphere (Arabic output), ViaVoice Translator (French/English, German/English, Italian/English, Spanish/English), Vocal PeTra (Italian/English). But none of this is *true* speech translation.

There are several other language technology applications where MT might soon be appearing as an additional facility. In the field of information retrieval there is much research at present on what is referred to as cross-language information retrieval (CLIR), i.e. information retrieval systems capable of searching databases in many different languages. Either they are systems which translate search terms from one language into other, and then do searching as a separate operation, with results presented en bloc to users; or, more ambitiously, translation of search terms and translation of output is conducted interactively with users. An area where MT is already involved is that of information filtering (often for intelligence work), i.e. the analysis of foreign language texts by humans. MT is used by many intelligence agencies in order to detect which documents are of interest and which merit full translation. A more recent area of interest is that of information extraction – extraction of pieces of information about individuals or companies or events from texts – performed by a large number of corporations as well as intelligence agencies. Currently systems for this task are monolingual (mainly English), but now there is the development of multilingual systems, where MT of some kind has a role – indeed, this is an area of intense activity within US government agencies, particularly Arabic into English.

As a final example, there is summarization. Most people when faced with a foreign language text do not necessarily want the whole text translated, what they want is a translated summary. There is a clear need for the production of summaries in languages other than the source. Summarization itself is a task which is difficult to automate (Mani 2001); but applying MT to the task as well is an obvious expansion, either by translating the text as a whole into another language and then summarising

it, or by summarising the original text and then translating the summary. The latter has usually been the approach of researchers so far. Working systems are some way in the future.

6. MT and its usefulness

When does MT work and when does it not work? Beyond the scope of MT is any prospect of fully automatic general-purpose systems capable of good quality translations without human intervention – and this is likely always to be so. In other words, we are not going to get MT systems that can take *any* text in *any* subject and produce *unaided* a good translation. Literature, philosophy, sociology, law and any other areas of interest which are highly culture-dependent are beyond the scope of unaided MT. It is true now, and will probably always be true.

In large corporations MT has proved to be cost effective – but only if the input is controlled, if terminology is standardized, if multilingual output required, if documentation is repetitive and if a restricted language is used for input texts – not all these conditions are necessary but most have to be present if the application is to be successful. When we come to personal MT use, where results do not have to be published, then obviously MT does work, and it is used – whenever the translation is wanted immediately and if there is no possibility or necessity of good quality revision of output. MT is also feasible for small-scale documents within a restricted domain, and if interactive assistance can be provided.

Why does MT fail? The reasons are much the same as for human translation. If the individual or the machine does not have sufficient knowledge of the source language then translation will not work. If the knowledge of the individual or the machine is not sufficient for the subject of the text then translation will fail – to varying degrees of seriousness. If the translator or the machine lacks specialized knowledge or lacks access to specialized lexica then there will probably be failure – again, to various degrees of seriousness. If the human translator has inadequate familiarity with the cultural background the translation will often fail; and MT must always fail wherever cultural background is needed because computers do not have any culture. If there is inadequate knowledge of, or data about, the target language in the relevant subject domain, then the translator or the machine will fail. If the human translator lacks experience in translation then the translation may not be adequate; the first use of a MT system may be inadequate for the same reason. Both can improve: the human translator with experience, the MT system from feedback incorporated either automatically (e.g. statistically) or via updating dictionaries and grammars by users or developers.

Has MT improved over the past 50 years? It depends on what is meant by improve. If we are talking about translation quality then perhaps not by much in the last 10 years (Hutchins, 2003). But if we are talking about the ease with which MT systems can be used, then there have been definite improvements. And if we are talking about the adaptability of systems to changed circumstances, to different types of equipment, then there have also been definite improvements. But these are all technical improvements. Linguistic improvements, improvements in the quality of translations, have been very slow – primarily because there are numerous problems in processing natural languages which are difficult to solve in practice, whatever the advances in our understanding of language and whatever the computer facilities being used. Many of these problems are inherent in language and linked closely to cultural differences. They will always be present and it is doubtful that any machine we can

envisage at present will ever be able to overcome them. This article has deliberately not covered research activity in the MT field. There are many developments which promise advances in the future, particularly in the areas of statistical and example-based methods.

Quality apart, there are various challenges for the future. Speech translation is a long-term aim: everyone would like to have it, but it will not come in the near future. There is a great need for systems for minor languages of Africa, Asia, India – as mentioned already, these languages are neglected by the commercial MT manufacturers. There is a need for MT systems for non-national languages within certain countries, such as Basque and Catalan in Spain, and Welsh in Britain – so far, there are no systems for these on the market. There is a definite need for systems designed for monolinguals, people who know nothing of either a source or target language. At present, current systems require people to know something of the language: if they do not know the source language to some extent they are often very difficult to use (e.g. people not knowing Japanese cannot enter Japanese text at a keyboard), and if they do not know the target language then they cannot judge whether the output is satisfactory or acceptable. Then, as already mentioned, there is the challenge of achieving the steady improvements that users expect in PC commercial software – they will expect similar improvements in MT software. We may expect increasing integration of MT in other PC software – word processors and web browsers – currently, MT software is a separate purchase, but there is no reason why translation should not offered as part of standard PC packages. Likewise we may expect MT components in software for summarization, information retrieval, information extraction, question-answering and other Internet applications. Finally, of course, there is the urgent need for consumer surveys, both of MT systems destined for the general public and of systems for corporate users.

It has to be stressed and kept constantly at the forefront of our minds that translation is a means and not an end. It is a means of facilitating cross-language communication. So too are MT and other computer-based translation tools. Perfect results are not crucial. Usability is more important. If publishable quality translations are required then human revision of MT output will always be necessary – as it is for most human translation. If MT output is used for assimilation or interchange functions, then perfectionism is by no means essential: imperfect communication can be tolerated in most circumstances. In sum, MT is a tool for use as and when required in order to save costs or effort in circumstances where cross-language communication is desirable or important – i.e. MT is a tool to aid bilingual communication rather than 'translation' in the strict sense. Whether MT should be used or not is a decision which should be based not on whether the system produces 'real' translations, not on whether it produces 'good' translations, but on whether the output can be used or is useful, and on whether use of the output can save time or money. An excessive insistence on top quality is a diversion from practical and useful application.

7. Sources of information

For general introductions to the practical application of MT and translation tools see the articles in Newton (1992) and Somers (2003a). For online documentation see the websites of the European Association for Machine Translation (www.eamt.org), the Association for Machine Translation in the Americas (www.amtaweb.org), *Translatum* (www.translatum.gr/dics/mt.htm), the *Machine Translation Archive* (www.mt-archive.info), and my own website (http://ourworld.compuserve.com/homepages/WJHutchins). The latest information

about MT activity (including research developments) will be found in the proceedings of the *MT Summit* conferences, and AMTA and EAMT conferences and workshops (see the AMTA and EAMT websites). For commercial aspects of the field see the Aslib conferences (www.aslib.com), the conferences held by LISA (www.lisa.com), and the journal *Multilingual Computing and Technology*.

8. References

Allen, J. (2003): 'Post-editing'. In: Somers (2003a), 297-317.

ALPAC (1966): *Languages and machines: computers in translation and linguistics*. A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council. Washington, D.C.: National Academy of Sciences – National Research Council.

Beaven, J. (1998): 'MT: 10 years of development', *Terminologie et Traduction* 1998: 1, 242-256.

Carl, M. and Way, A. eds. (2003): *Recent advances in example-based machine translation*. Dordrecht: Kluwer.

CLAW (1996): *Proceedings of the First International Workshop on Controlled Language Applications, 26-27 March 1996, Louvain, Belgium*. {Subsequent CLAW meetings held in 1998, Pittsburgh; 2000, Seattle; and 2003, Dublin – see EAMT-CLAW 2003).

Dodd, S.C. (1955): 'Model English'. In Locke, W.N. and Booth, A.D. (eds.), *Machine translation of languages* (Cambridge, Mass.: The MIT Press), 165-173.

EAMT-CLAW (2003): *Controlled language translation*. Proceedings of joint conference combining the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop, Dublin City University, 15th-17th May 2003. [Dublin: DCU.]

Esselink, B. (2000): *A practical guide to localization*. Rev.ed. Amsterdam: John Benjamins.

Esselink, B. (2003): 'Localisation and translation'. In: Somers (2003a), 67-86.

Gaspari, F. (2004): 'Online MT services and real users' needs: an empirical usability evaluation'. In: Frederking, R.E. and Taylor, K.B. (eds.), *Machine translation: from real users to research. 6th Conference of the Association for Machine Translation in the Americas, AMTA 2004*, Washington, DC, USA, September 28 – October 2, 2004. Proceedings. (Berlin: Springer Verlag, 2004), 74-85.

Henisz-Dostert, B. (1979): 'Users' evaluation of machine translation'. In: Henisz-Dostert, B., Macdonald, R.R. and Zarechnak, M. *Machine translation* (The Hague: Mouton), 147-244.

Hutchins, W.J. (1986): *Machine translation: past, present, future*. Chichester: Ellis Horwood.

Hutchins, J. (1998): 'The origins of the translator's workstation', *Machine Translation* 13 (4), 287-307.

Hutchins, J. (2001): 'Machine translation over fifty years'. *Histoire Epistémologie Langage* vol. 23 (1), 2001, 7-31.

Hutchins, J. (2003): 'Has machine translation improved?' *MT Summit IX: proceedings of the Ninth Machine Translation Summit, New Orleans, USA, September 23-27, 2003*. [East Stroudsburg, PA: AMTA, 2003] p.181-188. [A longer version is available on my website]

Hutchins, J. (2004): 'The Georgetown-IBM experiment demonstrated in January 1954'. In: Frederking, R.E. and Taylor, K.B. (eds.), *Machine translation: from real users to research*. *6th Conference of the Association for Machine Translation in the Americas, AMTA 2004*, Washington, DC, USA, September 28 – October 2, 2004. Proceedings. (Berlin: Springer Verlag, 2004), 102-114.

Hutchins, J., Hartmann, W. and Ito, E. (2004): *Compendium of translation software: commercial machine translation systems and computer-based translation support tools*. 9th edition, June 2004. [Available at http://www.eamt.org; new editions appear at approximately six month intervals.]

Kring, H.P. (2001): *Repairing texts: empirical investigations of machine translation post-editing processes*. Transl. by G.S.Koby et al. Kent, Ohio: Kent State Univ.P.

Kurematsu, A. and Morimoto, T. (1996): *Automatic speech translation: fundamental technology for future cross-language communications*. Amsterdam: Gordon & Breach.

LREC (2000): *Second International Conference on Language Resources and Evaluation, May-June 2000, Athens, Greece*. [Paris: ELRA]

Mani, I. (2001): *Automatic summarization*. Amsterdam: John Benjamins.

Melby, A.K. (1992): 'The translator workstation'. In: Newton (1992), 147-165.

Newton, J. ed. (1992): *Computers in translation: a practical appraisal*. London: Routledge.

Nyberg, E., Mitamura, T. and Huijsen, W.O. (2003): 'Controlled language for authoring and translation'. In Somers (2003a), 245-281.

Senez, D. (1995): 'Developments in Systran', *Aslib Proceedings* 47 (3), 99-107.

Somers, H. ed. (2003a): *Computers and translation: a translator's guide*. Amsterdam: John Benjamins.

Somers, H. (2003b): 'The translator's workstation'. In: Somers (2003a), 13-30.

Somers, H. (2003c): 'Translation memory systems'.In: Somers (2003a), 31-47.

Somers, H. (2003d): 'Sublanguage'. In: Somers (2003a), 283-295.

Sprung, R.C. ed. (2000) *Translating into success*. Amsterdam: John Benjamins. (American Translators Association Scholarly Monograph Series, 11)

Vasconcellos, M. (1986): 'Functional considerations in the postediting of machine-translated output', *Computers and Translation* 1 (1), 21-38.

Wagner, E. (1985): 'Rapid post-editing Systran'. In: Lawson, V. (ed.), *Tools for the trade: Translating and the Computer 5* (London: Aslib), 199-213.

Wahlster, W., ed. (2000): *Verbmobil: foundations of speech-to-speech translation*. Berlin: Springer.

White, J. (2003): 'How to evaluate machine translation', in: Somers (2003a), 211-244.

Woszczyna, M. et al. (1998): 'A modular approach to spoken language translation for large domains'. In: Farwell, D., Gerber, L. and Hovy, E. (eds), *Machine translation and the information soup: third conference of the Association for Machine Translation in the Americas, October 28-31, 1998, Langhorne, Pa.* (Berlin: Springer), 31-40.