Meta
Journal des traducteurs
Translators' Journal

Translation Tools

Lorna Balkan

Volume 37, Number 3, septembre 1992

URI: https://id.erudit.org/iderudit/001954ar
DOI: https://doi.org/10.7202/001954ar

See table of contents

Publisher(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (print)
1492-1421 (digital)

Explore this journal

Cite this article

Balkan, L. (1992). Translation Tools. *Meta*, *37*(3), 408–420.
https://doi.org/10.7202/001954ar

Article abstract

L'article fait état de l'influence de l'ordinateur sur le travail du traducteur et se penche sur les ressources que l'informatique aura bientôt à offrir dans ce domaine. On fait d'abord une distinction entre les différents types de systèmes de traduction puis on en décrit plusieurs afin de donner une idée générale des systèmes présentement disponibles et de ce que les futurs systèmes pourront offrir.

TRANSLATION TOOLS

LORNA BALKAN
University of Essex, Colchester, U.K.

Résumé
L'acticle fait état de l'influence de l'ordinateur sur le travail du traducteur et se penche sur les ressources que l'informatique aura bientôt à offrir dans ce domaine. On fait d'abord une distinction entre les différents types de systèmes de traduction puis on en décrit plusieurs afin de donner une idée générale des systèmes présentement disponibles et de ce que les futurs systèmes pourront offrir.

INTRODUCTION

This paper looks at the impact the computer has had on the work of translators, and how it may be of further assistance to them in the future. After an initial categorisation of the field I will go on to describe several systems in some detail. My aim is not to present an exhaustive list of systems, but rather to give a broad outline of what is currently available. Without the benefit of seeing demonstrations of the majority of the systems described, I have had to rely on secondhand sources and publicity material, so I cannot guarantee the accuracy of the information provided. Likewise I do not pretend to give a comprehensive list of all the areas of current research, but merely to give a flavour of what future systems might offer.

Note that I am concerned here only with the professional translator who either works at home on a freelance basis, or who is a member of a translation team in either a state or private organisation. I shall therefore not consider automated phrase books, and such like which are intended to be used primarily by non-linguists.

OVERVIEW

A distinction is generally made between machine translation (MT) systems and machine-assisted translation systems (MAT or CAT). Definitions vary. Slocum (1988) defines MT systems as those which are intended to perform translations without human intervention (excluding pre-editing or post-editing), while CAT is used to refer to systems where the human interacts with the computer at various defined stages during the translation process. According to his definition, CAT tools range in scope from word-processing and dictionary retrieval devices to systems that actually perform a translation. As with MT systems, post-editing is often appropriate. We intend to depart from this definition and use "machine translation" to refer to any system that actually performs a translation. This is because the definitions based on where and when the human interacts with the computer become confusing when, as we shall see, there are systems which operate either in interactive or batch mode. On the other hand, we continue to classify any other computerised translator tool which falls short of translating as a CAT device.

We shall look first at some CAT tools (1-4) and then at some machine translation systems (5):

1. text-processing tools;
2. the lexicon;

3. terminology;
4. integrated systems
   (where text-processing and lexical/terminological tools are combined);
5. machine translation systems.

TEXT-PROCESSING TOOLS

Translators, in common with some other professionals, are first and foremost in the business of processing texts. The fall in price of word processors over the last few years has brought them within the means of most translators. Equipped with a high quality printer, which has likewise come down in price, a freelance translator can revise and print his translation much more efficiently than by using a typewriter. In addition, even the cheapest word processors offer the facility of a spelling checker. Spelling checkers usually work by comparing each word of a text against a standard dictionary and highlighting those words which are not found. The user may then either change the spelling of the words in the text, or alternatively add the new words to the dictionary.

For a little more outlay a translator can invest in a personal computer, such as the IBM range of PCs. This provides him with a system for storing and filing texts. There is also a range of useful software available to him. For example, IBM produce a grammar checker called CRITIQUE (formerly EPISTLE — see Heidorn *et al.* 1982), which can detect some syntactic errors, such as wrong subject/verb agreement, (*we accepts your offer of employment) and incorrect verb forms (*we are informed you that), and some typing errors such as the duplication of articles, etc. The system works by undertaking a syntactic parse of the text and indicating the places where the parse fails.

Another useful facility on PCs is split screens. This enables the translator to consult the source and target texts simultaneously. This presupposes, of course, that the source text has been entered into the computer. The process of typing the text in can be very time-consuming and therefore not cost-effective. However, devices are now available for converting printed texts into machine-readable form, although all but the most expensive (in excess of 5,000 pounds) are unsuitable for converting non-standard characters and less than perfect paper copies into machine readable form.

THE LEXICON

(The discussion below centres on dictionaries, which are by no means the only reference works which a translator needs in his work (he often needs access to specialist journals, encyclopedias, etc.), but they are arguably the most important, since it has been estimated that conventional dictionary lookup can take up to 60 per cent of a translator's time. Much of what follows with respect to online dictionaries and databases applies equally to other forms of reference material).

The translator uses many kinds of dictionary including monolingual, bi- or multilingual, technical dictionaries or glossaries (see under terminology below), and thesauri. All printed dictionaries share the following disadvantages:

1. they are necessarily limited in size, so that the compiler has to be selective about what entries he includes, and what information to give on each entry;
2. language is dynamic, but the length of time and cost involved in producing a printed dictionary means that it will always be out of date;
3. most dictionaries are alphabetically ordered (the exception is thesauri which are conceptually ordered) which means that grammatical and semantic relations are obscured;
4. they are bad at representing multiword units — often a long search is required under several entries until the phrase is located.

The introduction of the computer to lexicography has had several consequences for the translator. First of all, many dictionaries are now available in machine readable form (e.g. *Collins bilingual dictionaries*). This means that the translator who is equipped with a computer can consult them on-line, without having to leaf through paper dictionaries, thus saving valuable time.

Secondly, the exploitation of database management techniques in lexicography promises to revolutionise both dictionary production and consultation. For a discussion of lexical databases see Bennett *et al.* (1986). Briefly, a lexical database, which corresponds to a lexical entry in a dictionary, consists of a structured collection of records. Records are subdivided into subfields which contain information about register, definition, etc. Unlike a conventional dictionary, where, as we noted, relationships between entries are difficult to represent, relationships between records and subfields are represented explicitly in a database.

Databases can cope with the dynamism of language better than printed dictionaries, since entries are constantly being added to and modified. Therefore a translator who has access to a database always has up-to-date information. Furthermore he can access this information in multiple ways, since he can search the data along different dimensions (e.g. he can search for words according to morphological, semantic, etc. criteria) and in any number of combinations.

Even translators who have no access to a computer can benefit indirectly from lexical databases, since databases permit revised editions of dictionaries to be published more frequently and at reduced cost, and allow a greater variety of product, since in theory any subset of the database may be published as a dictionary.

Another important aspect of databases is that they can potentially be used by both humans and machines, so can be used as the basis for a variety of natural language processing systems. A database for *The Oxford English Dictionary* is currently being developed at Waterloo and it is anticipated that it will have a wide range of applications.

TERMINOLOGY

Since many translators work in specialised areas, the use of computerised terminological resources is potentially important. A term data bank (or "termbank") is used to refer to a collection of databases containing terminological data. The functions of a termbank are similar to that of a lexicographical database, with the additional role of controlling neologisms and harmonising existing usage.

A recent survey of translation practices (Smith and Tyldesley 1986) shows that only a small percentage (14%) of translators actually use a termbank. Some firms have inhouse termbanks. One example is Siemens, whose termbank TEAM was set up in 1967. Siemens also makes its termbank available to partner companies and other external subscribers in return for supplying terminology. Eurodicautom is the name of the termbank of the EEC, and was created in 1973. It appears to be the most popular termbank with translators who have access to a termbank, but its popularity may stem from the fact that it is the most widely available online termbank.

The fact remains that most termbanks in the world are not accessible to translators, at least online. Even if they are available, information retrieval from them may be hampered because the user must be familiar with their characteristics and query language, which vary from termbank to termbank. It may therefore not be cost-effective for translators to carry out their own searches. Some organisations with inhouse term banks require searches to be carried out by trained personnel on behalf of the translators. But for the outside translator search charges may also prove quite costly. One exception is Termium, the termbank of the Canadian government, which offers online search free of charge (excluding tele-communication charges).

As with dictionaries, the data in termbanks can be made available to a wider audience through the publication of paper glossaries.

INTEGRATED SYSTEMS

A picture is beginning to emerge of what computer tools the translator needs. Ideally he will have access to termbanks and lexical databases. Otherwise he will have online dictionaries and glossaries derived from them. He will also need software packages to construct and manage his own glossaries. What is available commercially to help him in these tasks?

LinguaTech's TERMEX is a multilingual management system, which allows the user to consult machine readable dictionaries, and cut out and "paste" the appropriate entry into a text without leaving the word processing environment. TERMEX glossaries can be purchased with the system, or constructed by the user. Facilities are available for combining glossaries, extracting subsets and converting glossary files to and from standard wordprocessing files.

A slightly more sophisticated system is offered by INK TextTools. It performs the same tasks as TERMEX, but provides additional facilities for translators who have the source text in machine readable form. It facilitates the construction of specialised terminology dictionaries based on the actual texts being translated by analysing the source text and automatically compiling word lists and bilingual glossaries. At the same time it provides frequency and context information. It also offers a morphological component, available as yet for English only, where words in the source text are analysed to their morphological roots, and a sorted list of possible compounds (eg. "computer programming language") produced.

None of the translation aids described so far actually perform translations: they are principally concerned with terminology and word processing. Nevertheless, they are clearly of benefit to the translator. Melby (1987a), estimates that a translator, even without having the source text in machine readable form, but who is equipped with a microcomputer with multilingual word processing software (possibly with spelling checkers, word counting, grammar checkers and synonym finding), a multilingual terminology management system of the TERMEX type and a desktop publishing system can increase his productivity by 40 per cent.

MACHINE TRANSLATION

The next step in automation is systems that do actually produce draft translations. Until recently the main commercial systems were ALPS, Weidner, LOGOS and SYSTRAN. Before continuing, it is useful to draw up a list of criteria by which to judge translation systems:

1. Cost-effectiveness: this is affected by such things as initial investment and whether the source text is in machine-readable form. It is also affected crucially by the rate at which the system translates and the amount of post-editing (or pre-editing) which it requires, and whether the text would be revised anyway even if it were translated by a human. Note that depending on the customer for which the translation is intended, an unedited piece of machine translation may be acceptable, as long as it is intelligible and faithful to the source text.

2. User-friendliness: in the case of interactive systems, this includes the way in which the user interacts with the system (e.g. is the system menu-driven?), and whether the translation system is fully integrated into a word processing environment.
3. Versatility of the system: what language pairs can it translate, and what subject areas? Does it supply any glossaries of its own?
5. The depth of linguistic analysis which the system performs, and its ability to treat such linguistic phenomena as idioms and compounds.
6. The robustness of the system: what does it do when it encounters unfamiliar words or syntactic constructions, and how does it cope with ambiguity (does it, for example, choose one solution only?).

Neither of the first two systems we will look at, ALPS and Weidner, claim to offer any substantial degree of machine translation.

ALPS

ALPS is the translation support system developed by Automated Language Processing Systems, Inc. (ALPS), a private company set up in 1980 (ALPS has since changed its name to ALPNET). The ALPS system is intended as a set of translator tools, with the stress on interactivity between the system and the translator at each stage of the translation process. The software runs on the IBM PC/AT and IBM 4300 series, and the translation pairs offered are currently from English to French, German, Spanish, Italian; French to English and German to English.

At one end of the range ALPS can be used as a sophisticated multilingual word-processor and dictionary management system. Like TextTools it can conduct an automatic search of online dictionaries and display the translations for source text words. Dictionary entries can be incorporated into the translation text by a simple key stroke.

At the highest level the ALPS system can be used to make draft translations, which are produced sentence by sentence, so that the translator can modify them as he goes along (the source and target text are displayed side by side on the screen). The system first examines the source text sentence by sentence and creates a list of words that are not found in the dictionaries. These words can be entered, together with their translations and their word class into dictionaries. The word class is used by the "grammatical processing" stage that follows. The entry for the word "back", for example, would look like:

French
back
dos [n]
soutenir [v]
arrière [adj].

Where the system encounters lexical or grammatical ambiguity, it turns to the translator for help. For example, "boy's" in

"did you believe the boy's broken promise"

can have two interpretations: it is either a possessive marker or a contracted form of "has". The system further relies on the translator to reorder sentences where the construction of the target language does not mirror that of the source language. In short, interaction with the translator is used to compensate for the poor level of linguistic processing which the system provides.

WEIDNER

Established in 1977, Weidner offers two translation systems: MacroCAT, which runs on a mainframe computer, and MicroCAT, which was the first PC based translation system to appear on the market. Language pairs include English to French, German, Italian, Spanish and Portuguese and French to English and Spanish to English. Like ALPS, the Weidner system is embedded within a sophisticated word processing environment, and is menu-driven.

Prior to translation the system conducts a vocabulary search and produces a list of not-found-words, which it invites the user to enter into the dictionary. From then on, the system is no longer interactive. Like ALPS the Weidner system relies heavily on information contained in the dictionary. However, dictionary entries contain more information than their ALPS counterparts, and include details of inflection and the syntactic behaviour of individual words and idioms. There is also an attempt to introduce semantic information with features such as "human, group, body part, animal, inanimate, concrete and abstract" for noun entries. This information is used in the analysis part of the system to build a phrase structure tree. Analysis is fairly shallow and consists of idiom search, homograph resolution and structural analysis, all of which is performed on a local level only (at the noun phrase and verb phrase level only). Ananiadou (1987) reports that raw output from Weidner suggests that the homograph resolution is not always successful. Also, as in ALPS there is no attempt in the later stages of processing to reorder the text to suit the target language structure.

The system is supplied with a core dictionary of between 9,000 to 15,000 words and idioms, but the translator is expected to create his own specific dictionaries.

LOGOS

The LOGOS translation system of LOGOS Corporation is a slightly more sophisticated system than either ALPS or Weidner. Although it has been in development since 1969, it was first made public in 1982. Unlike the Weidner and ALPS systems, LOGOS does not actually sell its system, but charges monthly on-site leasing fees, and runs on a mainframe. The language pairs currently offered or under development are German to English and French, and English to German, French and Spanish.

In common with the previous two system, LOGOS is embedded within a powerful word processing system, and is heavily dictionary-based. It is menu-driven and like the Weidner it is interactive only up to the dictionary update stage. From then on it operates in batch mode. Analysis is much more semantically driven than the other two systems, the linguistics being based on a simplified case system. The user is required to specify correspondingly more semantic information in the dictionary entries. He can also enter translations for combinations of words, as well as for single words by writing so-called "semantic rules". He can for example ensure that the "treffen" in the German expression "*Entscheidung treffen*" (take a decision) is translated as "take" and not literally as "meet".

Another novel aspect of the system is that compound words encountered in the not-found list are given a suggested translation, which is just the translation of their component parts. These translations may be accepted or rejected by the translator at will. In addition, the translator can also steer the translation towards a particular subject field, by specifying up to five subject fields before translation commences (the system recognises up to 246 different subject fields).

Translation speeds of 1000 words per hour have been reported (Ananiadou 1987). LOGOS is used by several multinationals, including Hewlett-Packard, and by the Canadian government which has chosen LOGOS as the general purpose machine translation component in their network-based automation of their translation services.

SYSTRAN

Developed in the 1960's, SYSTRAN was bought by the Commission of the European Communities in 1975, and is currently available in English to Italian, Arabic and Russian, and from French to English, German to English and Russian to English. It runs on a mainframe computer in batch mode only. Its linguistic analysis is similar in depth to that of LOGOS. Interesting features include an attempt to guess at the translation of some unknown words, based on morphological rules. For example, in the French-English module, the ending of "*radiologue*" would be stripped, and the corresponding English ending added, giving in this case the correct translation "radiologist".

SYSTRAN continues to be used in the EEC, and has recently been made available to the general public in France through Minitel. It is also offered via network access to users in North America. SYSTRAN is capable of translating up to 150,000 words per hour (Wheeler 1987).

ASSESSMENT

Given the limitations of these systems, it is instructive to ask how useful they are to the translator. A recent conference on machine translation (Mayorcas 1990) found that the cost-effectiveness of machine translation depends crucially on non-linguistic factors, such as whether or not the source text is in machine readable form. This is borne out by Ingall (1988?) who reports that Berlitz Translation Service, the largest translating firm in the USA bought ALPS for inhouse use but stopped using it partly because of the cost and time needed to enter text into the computer.

The advantage of the above mentioned translation systems is that they are able to produce translations within a reasonable processing time, even if the standard of translation is not always very high. In general, machine translation systems lend themselves better to the translation of technical texts, where the range of vocabulary and even grammatical expression is restricted.

The theme of restricting text type in order to improve the performance of a machine translation system is found most notably in METEO, the system developed by the TAUM group to translate weather bulletins from the Canadian Meteorological Office from English into French. Not only is the vocabulary restricted to the sub-language of meteorology, but the bulletins themselves are written in a highly telegraphic style, where sentences usually lack a tensed verb. The success rate is over 80 per cent, and these 80 per cent do not need any human post-editing (Isabelle 1987). Another system which uses limited syntax and vocabulary is TITUS, which translates abstracts of articles in textile terminology for the *Institut textile de France*. The abstracts are in fact rewritten by a human, to ensure that they are in a form appropriate for the translation system. This is known as pre-editing.

Both SYSTRAN and the Weidner system have been used successfully with restricted syntax documents by Rank Xerox and by Perkins Engineering Ltd, UK respectively. A description of the restricted language which Perkins uses, Perkins Approved Clear English (PACE), is given in Pym (1990). It was devised for a dual purpose: to make instructions in the company's manuals clear or clearer to non-native speakers of English, and to aid translation, whether carried out by traditional or computer-aided methods. Apart from the simplified syntax, only one sense is assigned to each word, so that lexical ambiguity never arises. Using PACE with the Weidner system has cut translation costs, Pym estimates, by 50 to 70 per cent.

However it is not usually viable to pre-edit a text or restrict text types to sub-languages. In the case of PACE and TITUS, pre-editing is an integral part of the document creation process anyway. The upshot is that smaller systems such as ALPNET are failing to attract enough customers and are having to diversify their business interests.

Alps has acquired a number of translation agencies and has changed their emphasis from selling translation software to selling translations produced using the software. Customers who are satisfied with the product may then acquire the software for themselves. Zirkle, the president of ALPNET London, points out (Zirkle 1990) that the total revenue from all software sales over the past ten years would not equal half of the money spent by CAT sellers in development and marketing over the same period. Weidner similarly has experienced financial difficulties. The company has now been taken over by Bravice, one of its Japanese customers.

What then are the prospects for a general purpose machine translation system in the foreseeable future?

The systems described above all belong to the first generation of machine translation systems. Several characteristics distinguish them. First of all, they are "direct", which is to say that the translation rules operate directly on the input string, rather than on a representation (e.g. a syntactic tree) of the input string. The emphasis is on dictionary coding and on *ad hoc* rules. Secondly, in first generation systems there is typically no separation of algorithms and linguistic data. This makes them difficult to maintain and upgrade, since the more rules that are added to the system, the more opaque their overall behaviour becomes.

A second generation of machine translation systems has since emerged which incorporate three main design features. Firstly, they are "indirect" in that translation rules apply not to the input text directly, but to a representation of the input text, produced by the analysis phase. The target text may be generated directly from this representation, in which case the system is said to be "interlingual", or alternatively, the source text representation may be mapped on to a target text representation, from which the target text is then generated (the so-called "target" approach). Second generation systems typically adopt the target scheme. Thirdly, second generation systems are characterised by a separation of linguistic data (described as a set of grammar rules) and the algorithms that operate on the linguistic data, i.e. interpret the rules). Grammars written in this way are much more perspicuous and easy to modify. System maintenance is further facilitated by keeping analysis and synthesis strictly monolingual. This goal, only partially achieved in systems like SYSTRAN, also means that the same analysis and synthesis modules can be reused for any number of language pairs.

METAL

So far, few machine translation systems based on new technology have found their way on to the market. METAL, the machine translation system developed by Siemens, became available in 1988 and claims to be the first commercial machine translation system to make "full use" of artificial intelligence (AI). Compared to earlier systems, it employs some interesting linguistic and computational techniques. It is a transfer system based on a context free grammar which uses case frames to help identify a predicate and its arguments. A syntactic tree representation is built during the analysis phase. The grammar rules of the analysis module are grouped into different levels. The bottom level rules are most likely to be relevant, and are applied first. Only if they fail to apply does the system try to apply higher level rules. Thus, it is claimed, the "most likely" reading of ambiguous sentences is obtained.

METAL also uses a menu-driven "expert system" to update its dictionaries. The user need only enter the word and its category and the system automatically encodes almost all the features and values of the new entry, based on morphological analysis of the root form and by comparing it with similar entries already in the dictionary. The user may then review or revise the default entry at will.

Like LOGOS, METAL offers default translations for unrecognised compounds, which are claimed to be correct in 70 per cent of cases. Another key feature of METAL is its preservation of the layout of the input text. It automatically deformats the source text and reformats the target text. This is important, because it has been estimated that less than 60 per cent of each page of a technical document is translatable material, the rest consisting of charts, diagrams, etc.

METAL observes a certain amount of modularity, in that the analysis module is claimed to be target language independent. It is currently available in German to English and English to German, with pairs with French, Spanish and Dutch being planned and is used by several government and educational establishments in Germany and Holland. Translation speeds are given as one word per second. The price of the system is $131,500.

TOVNA

Another newcomer, TOVNA machine translation system from Israel, also claims to use AI techniques and to be the first translation system that "learns". This claim is based on the fact that not only can the user add phrases and words to the lexicon in the usual way, but the system can also derive phrase structure rules from these phrases. "Learning" is mostly done through an interactive process of "diagnosis", teaching and retranslation (see Bedard 1990). The system keeps a record of each step it takes in the processing of a sentence. If at the post-editing stage the user discovers an error, he can consult this record of the translation to find out the cause. For example, if the English phrase "the liquid oxygen tank" is wrongly translated into French as "le réservoir liquide d'oxygène", the user can consult the parse tree (represented in parentheses thus: "(the (liquid (oxygen tank))") and alter it to "the ((liquid oxygen) tank)". The system tends to "remember" that ((ADJ + N1) + N2) is more frequent than (ADJ + (N1 + N2)), and will eventually alter its default rules accordingly. Likewise, the system infers that liquid goes well with oxygen and other semantically similar words and will give preference to future parses which have the same semantic relationships. Here, the notion of preference based on statistics is important.

Diagnosis also enables the user to examine which patterns from the bilingual phrase table have applied during transfer. He can then choose another pattern or make a new one by entering a typical word sequence as an example. At the same time he can determine the degree of specificity of the new pattern. For instance, the following example could be entered: "Sixty million US dollars/*Soixante millions de dollars US*" together with the information that the rule is restricted to the word "million" preceded by "sixty" or any other word of the same word class, and followed by "dollars" or other words with the semantic feature "currency".

After "teaching" the system, the user can retranslate the sentence, and if necessary repeat the above processes until he is satisfied with the translation. The various functions of the system are menu-driven. Bedard points out that while this teaching facility gives the user some desirable control, it does require a certain degree of skill.

Despite these innovative features, TOVNA seems to perform fairly crude linguistic analysis confined to phrase level. Semantics appears to encompass little more than selectional restrictions. Nevertheless, TOVNA does claim to adhere to the principles of separation and modularity since it "separates the software from the language base" and "Once a language has been installed, it works with any of the other languages that exist in the database" (TOVNA publicity material). The translation can be performed either on batch or section by section (so-called interactively). In batch mode the system can translate around 3600 words per hour. Language pairs currently available include English to French and Russian. The system runs on mini computers and a perpetual licence costs $150,000 per language pair.

TOVNA is currently being used on an experimental basis by the World Bank in Washington, who report that results are "mixed but encouraging" (Bedard 1990).

GLOBALINK

The last few years have also seen the emergence of two other machine translation systems, both of which are PC-based and relatively cheap (c.$1,000). The first of these is Globalink, an American system from Hadron Inc., which was launched in 1989. Its syntactic parsing is no more sophisticated than SYSTRAN's, and it uses no semantic codes whatsoever. The lexical update procedure allows one to add new words and phrases to the dictionary, but otherwise offers no special features, such as default coding. Translation can be performed interactively sentence by sentence or in batch mode, where up to 20,000 words can be translated per hour.

The Globalink system is available in the following language pairs: English to French, Spanish, Chinese and Arabic; French to English, Spanish to English, Italian to English, German to English and Russian to English. In addition to the main dictionary, micro dictionaries are sold in specialist fields such as finance, computing, and law. Reviewed in *Electric Word* lately, the system was reported to give as good a translation as the Weidner MicroCat at a fraction of the cost.

LINGUISTIC PRODUCTS

Another new low-budget PC machine translation system is the Texas-based Linguistic Products. It is also a relatively simple system which is 80 per cent dictionary based. However, it has been used by the Madrid subsidiary of the Digital Equipment Corporation which claims that it is 70 per cent accurate in translating manuals. It has also been tested by Ian Piggott of the CEC's machine translation section who reports that it offers a translation level practically as high as that of Weidner, SYSTRAN or LOGOS, but at much less cost.

FUTURE TRENDS

It is fair to say that none of these systems radically improves the quality of raw machine translation output. Let us review the achievements of first and second generation systems. The strength of first generation systems is their strong lexical bias. Second generation systems on the other hand tend to emphasise syntax and suffer from a relatively weak semantic component. Most of these systems have a tree as their linguistic representation, and perform tree-to-tree transformations as their basic operation. Semantic information they use is generally restricted to selectional restrictions, predicate-argument structures and semantic roles. It is generally accepted that "third generation" systems will have to build more powerful mechanisms for semantic, textual and pragmatic analysis.

A number of Japanese firms (including Toshiba, Sharp and Hitachi, to name but a few), have recently launched commercial machine translation systems which claim to be semantically based, or even interlingual. Interlingual systems are not new — TITUS for example uses an interlingual model, but operates, as we have seen, with a restricted sub-language. However, there are many problems associated with interlingual systems. For example, how does one generate the target language when lexical and syntactic information about the source text has been neutralised into an abstract interlingual representation? It is interesting to see how successful these new systems will be.

Other areas of machine translation research, where no commercial systems are yet available, include systems based on constraint grammar formalisms such as LFG (Bresnan 1982) which use feature structures as their linguistic representations. Different kinds of information (syntactic, semantic, etc.) can be combined in one feature structure.

Investigation is likewise currently being undertaken on the use of formal semantics in machine translation (e.g. the Philip's Rosetta system) and on the application of statistics (e.g. work at IBM Yorktown heights and Buro voor Systemontwikkeling, Holland).

While on the one hand work continues on improving the theoretical foundations of machine translation, other researchers are concentrating on ways of making existing systems more user friendly. It is the latter line of research that is more likely to be of benefit to translators in the short term.

Alan Melby and his team at Brigham Young University in the United States have designed a machine translation system which aims to overcome the frustration frequently felt by translators at having to revise entire texts of raw machine translation, where the quality is inevitably uneven. The key notion is that of "selective translation"; the translator is not obliged to use the raw machine translation for every sentence of a text if it does not meet a particular standard. The system operates on three levels, all of which are fully integrated. The first level functions as a wordprocessor with online access to bilingual dictionaries, and possibly network access to databases of documents and terminology banks. The next level assumes that the source text is in machine readable form and enables the translator, among other things, to "point" at a word in the source text and automatically see its translation equivalents from the dictionaries. On the third level the machine translation is performed, but on a selective basis. Melby describes a selective machine translation system as follows:

> In this approach, the MT system grades its own output according to problems encountered during lexical, syntactic, or semantic processing. Then, the translator translates the text segment by segment until a segment is encountered for which there is a translation that received a grade above the translator's chosen threshold. Such a segment is likely to be worth post-editing and the translator evaluates it and inserts it into the text or ignores it and translates the segment. (Melby 1987b)

Otherwise the translator can fall back on the other two levels of the system. Melby believes that it should be possible to take existing machine translation systems and modify them to be selective translation systems.

Interactive systems are generally seen as one way of compensating for the inadequate linguistic and real world knowledge that we are at present able to build into machine translation systems. The place where interaction should take place (either pre-editing, post-editing or during the translation process) and the form the interaction should take, is a matter of some debate (see Melby 1987b for a discussion). Two research projects, the Knowledge Based machine translation project of Carnegie Mellon University (see Carbonell *et al.* 1987), and the Distributed Language Translation Project of BSO in the Netherlands (see Witkam 1983) are both pursuing the approach of interactive pre-editing. The advantage of pre-editing is that it can be performed by a user who only knows the source language. Carbonell *et al.* (1987) describe how a system can question the user to resolve three kinds of ambiguity: word sense ambiguity,

e.g. The word "pen" means:
1. a writing pen
2. a play pen
NUMBER?>

referential ambiguity

e.g. The word "she" refers to:
1. Cathy
2. my mother
3. the sailboat
NUMBER?>

and syntactic ambiguity

e.g. I saw a man with a telescope
1. "a man" and "a woman"
2. "a man" and "a woman with a telescope"
NUMBER?>
1. the action "saw" takes place "with a telescope"
2. "a man and a woman" are "with a telescope".
NUMBER?>

The user selects the correct interpretation by inserting the appropriate number. While parse trees could conceivably be used to illustrate syntactic ambiguity, the approach adopted above means that the user does not need any specialist linguistic knowledge.

Pre-editing is also advocated in the UMIST English-Japanese MT project (see Johnson *et al.* 1987). They are motivated by their belief that important insights into the organisation of machine translation systems can be obtained by studying the behaviour of the human translator. They start from the assumption that "a human translator is, first and foremost, a target language expert" (Johnson *et al.* 1987), and therefore suggest that if a machine is to behave functionally like a human translator, it has to be (amongst other things) a target language expert. On the other hand, they anticipate that the machine will be more or less deficient in the knowledge of the user's source language, a deficiency that can be remedied by consultation with a source language monolingual operator. They note, however, that since some apparent ambiguities in analysis may carry over to the target language, that some interaction should be handled in transfer.

Viewing machine translation systems as expert systems promises to be a fruitful line of research. An expert system is "a computing system capable of representing and reasoning about some knowledge-rich domain...with a view to solving problems and giving advice...it must be capable of explaining and justifying solutions and recommendations in order to convince the user that its reasoning is, in fact, correct" (Jackson 1986). In the case of machine translation, the system can model the translator's knowledge and thinking process. Zirkle (1990) deplores the fact that in CAT "translators and CAT developers still remain practically strangers to each other", a statement that could be applied equally to machine translation in general. However, as Sager (1990) observes: "It is now widely accepted that translation systems have to be based on models of translation which themselves require a sound theoretical foundation" but adds "It will take a few more years before this realisation is fully reflected in operational systems."

In conclusion, while sophisticated wordprocessing software and lexical aids are already proving their worth to translators, machine translation systems as such are still crude, and useful only in restricted environments, particularly sublanguages. Research into sublanguages, linguistic theory (especially in the field of semantics, discourse and pragmatics), the theory of translation, AI techniques (such as expert system design), and human-computer interaction should all contribute to more efficient and user friendly machine translation systems in the future.

BIBLIOGRAPHY

ANANIADOU, S. (1987): "A Brief Survey of some Current Operational Systems", *Machine Translation Today*, Edinburgh University Press, Edinburgh.

BEDARD, C. (1990): "TOVNA, a New Word for Teachable Machine Translation", *Electric Word*, 18.

BENNET, P.A., JOHNSON, R.L., MCNAUGHT, J., PUGH, J.M., SAGER, J.C., SOMERS, H.L. (1986): *Multilingual Aspects of Information Technology*, Gower Publishing Company Limited, Aldershot, Hants.

BRESNAN, J. (1982): *The Mental Representation of Grammatical Relations*, MIT Press, Cambridge, Mass.

CARBONELL, J. and M. TOMITA (1987): "Knowledge-based Machine Translation, the CMU Approach", *Machine Translation*, CUP.

HEIDORN, G., JENSEN, K., MILLER, L., BYRD, R., CHODOROW, M. (1982): "The EPISTLE Text-critiquing System", *IBM Systems Journal*, 21-3.

HUTCHINS, W.J. (1986): *Machine Translation: Past, Present, Future*, Ellis Horwood / Wiley, Chichester / New York, 1986.

INGALL, Martin (1988?): *Computer-aided Translation An Industry Survey*, Harvard University.

ISABELLE, P. (1987): "Machine Translation at the TAUM Group", *Machine Translation Today*, Edinburgh University Press, Edinburgh.

JACKSON, Peter (1986): *Introduction to Expert Systems*, Addison-Wesley.

JOHNSON, R. and WHITELOCK, P. (1987): "Machine Translation as an Expert System", *Machine Translation*, CUP.

KING, M. (1990): "Taking Stock: Machine Translation in the Last Five Years", *IES News*, issue No. 24.

KNOWLES, F.E. (1990): "Machine Aids for Translators: What does the Future Betoken?", *Translating and the Computer 10*, Aslib, London.

MAYORCAS, P. (1990): *Translating and the Computer 10*, Aslib, London.

MELBY, Alan (1984): "Machine Translation with Post Editing versus A Three-level Integrated Translator Aid System", *Proceedings of the International Seminar*, Cranfield Institute of Technology.

MELBY, Alan (1987): "Creating an Environment for the Translator", *Machine Translation Today*, Edinburgh University Press, Edinburgh.

MELBY, Alan (1987a): "The Translator Workstation", *Language and Technology*, vol. 2, Amsterdam.

MELBY, Alan (1987b): "On Human-machine Interaction in Translation", *Machine Translation*, CUP

MCNAUGHT, J. (1987b): "A Survey of Termbanks Worldwide", *Translating and the Computer 9*, Aslib, London.

PYM, P.J. (1990): "Pre-editing and the Use of Simplified Writing for MT: An Engineer's Experience of Operating an MT System", *Translating and the Computer 10*, Aslib, London.

SAGER, J. (?): "Ten Years of Machine Translation Design and Application: From FAHQT to Realism", *Translating and the Computer 10*, Aslib, London.

SLOCUM, J. with BENNETT, W.S., BEAR, J., MORGAN, M. and R. ROOT (1987): "METAL: The LRC Machine Translation System", *Machine Translation Today*, Edinburgh University Press, Edinburgh.

SLOCUM, J. (1988): *Machine Translation Systems*, CUP, Cambridge.

SMITH, D. and D. TYLDESLEY (1986): *Translation Practices Report*, Reading: Digital Equipment Corporation.

WARWICK, S. (1987): "An Overview of Post-ALPAC Developments", *Machine Translation Today*, Edinburgh University Press, Edinburgh.

WHEELER, P. (1986): "The LOGOS Translation System", *Proceedings of IAI-MT86*, Burgerhaus, Dudweiler.

WHEELER, P. (1987): "SYSTRAN", *Machine Translation Today*, Edinburgh University Press, Edinburgh.

WITKAM, A.P.M. (1983): *Distributed Language Translation: Feasibility Study of a Multilingual Facility for Videotex Information Networks*, Utrecht:BSO.

ZIRKLE, A.T. (1990): "The Role of Computer-aided Translation in Translation Services", *Translating and the Computer 10*, Aslib, London.

Journals include:

Electric Word, (Formely, "Language Technology" and "Language Monthly"), P.O. Box 70486, 1007 KL Amsterdam, the Netherlands.