WHY COMPUTERS DO NOT TRANSLATE BETTER

by W.John Hutchins
(University of East Anglia, Norwich, England)

[Paper for Translating & the Computer 13, London, November 1991]

Abstract

The linguistic and computational complexities of machine translation are not always apparent to all users or potential purchasers of systems. As a consequence, they are sometimes unable to distinguish between the failings of particular systems and the problems which the best system would have. In this presentation I shall attempt to outline the difficulties encountered by computers in translating from one natural language into another. This is an introductory presentation for those unfamiliar with what computers can and cannot achieve in this field.

Introduction

My intention is this paper is to provide some explanation for the difficulties encountered by present computer system which attempt to produce partial or complete translations of texts from one natural language into another. The emphasis will be on what can or cannot be achieved automatically at present.

I shall not be concerned with the relative merits of different approaches to translation problems, for example, whether systems which switch between languages through some kind of interlingual representation are better than those which do not, or whether systems which employ methods from artificial intelligence are better than those which are use more familiar methods of computational linguistics, and I shall say virtually nothing about what developments may bring improvements in the future. Furthermore, I shall not be describing any particular system of whatever kind, past or present, or any methods of analysis or processing, or how dictionaries may be structured and compiled, whether monolingual or bilingual.

My aim is to give an introduction, for those unfamiliar with machine translation (MT), to the main areas which must be taken into consideration, even if designers of particular systems have opted deliberately to ignore some of them. The aim is to highlight in the broadest terms those areas of translation which are relatively easy for computerised handling and those areas which are relatively difficult, and I am describing the present situation and make no predictions about the future. The purpose, therefore, is not to describe the inherent limitations of machine translation, but to give a rather simplified explanation of what can be expected from any system at the present time, whatever its particular methodology. (For more details about the way MT systems work see Hutchins & Somers 1992).

Finally, it should be clear that I am not providing a methodology for evaluating system, only giving at best a check list of areas in which evaluation can take place. Evaluation involves much more than the quality of translation, although that is obviously a most important aspect. It involves also, for example, the integration of an MT system in the whole processing framework: transmission and receipt of texts, formatting, dictionary updating, editing, printing, and distribution. It involves examination of the compatibility of systems with other computer facilities, and in particular it embraces the integration of the system into the working patterns, practices and attitudes of existing staff, and the aims of the organisation as a whole.

System types

Although my object is not to describe the different types of systems available, a brief outline is in order. In broad terms, we may distinguish firstly between systems which attempt to translate whole texts without intervention by human operators and those which require assistance to resolve problems of ambiguity in the source text or to select the most appropriate word or phrase in the target language. The former type – sometimes referred to as ‘batch’ systems because the whole text is processed as one task – generally produce output requiring revision by a human translator to a greater or lesser extent. This revision (or ‘post-editing’) may well be substantial if the text is intended for publication, but it may be minimal – or even absent – if the text is intended only for information-scanning purposes within an organisation. An alternative mode of operation with such systems is to ‘pre-edit’ the input texts to reduce ambiguities and complexities of structure. As a further step texts can themselves be composed in a ‘controlled language’, in which words are as far as possible used in a single meaning only and in which sentence structures are kept to the simple forms which the computer programs are able to handle. In ‘interactive’ systems the computer seeks the assistance of a human operator during the translation process itself in order to resolve ambiguities and make decisions which may be difficult for programmers to define rigorously.

The question to be asked is therefore why some problems are more difficult for computers to deal with than others? With this knowledge, users should be able to understand why when 'post-editing’ certain types of ‘mistakes’ need to be constantly corrected, why when ‘pre-editing’ texts or composing in controlled languages’ certain types of ambiguity and constructions must always be avoided, and why in ‘interactive’ systems certain types of questions recur again and again.

The methods for dealing with translation difficulties vary from system to system. In many cases, the ambiguities specific to the source language are tackled in operations separate from the treatment of differences between languages. Commonly three basic operations are recognised: the analysis of the source text, the bilingual transfer of lexical items and structures and the generation of the target text. Questions of ambiguity and choice occur at every stage. For example, resolving the ambiguity of English *cry* between ‘weep’ and ‘shout’ would be part of a program for the **analysis** of English. On the other hand, the selection of *connaître* or *savoir* in French for the English verb *know* would be a matter for a separate **transfer** program. Analysis involves also the identification and disambiguation of structures, e.g. whether *He saw her shaking hands* means that he saw someone who was welcoming a visitor or he saw someone who was suffering from the cold weather. Transfer likewise can involve changes of structure, e.g. from an English infinitival construction *He likes to swim* to a German adverbial construction *Er schwimmt gern*. **Generation** is often incorporated in transfer operations, but when a separate component it might include operations to distinguish between English big, large and *great* (about which more later) and the production of correct morphology and word order in the target language (*ses mains tremblantes*, *er darf nicht schwimmen*).

	Lexicon	Structure
Analysis:	cry→ weep/shout	morph: laughs/laughing syntax: he saw her shaking hands
Transfer :	know→ savoir/connaître	he likes to swim er schwimmt gern
Generation:	[grand]→ big/large	ses mains tremblantes er darf nicht schwimmen

In some earlier systems and still in a number of present-day microcomputer systems, all these operations are incorporated into single massive programs. Obviously, such systems – often called ‘direct translation’ systems – can only be designed for two particular languages and in only one particular direction. Many present systems, particularly if intended to be multilingual, have separate programs for each of these components; a procedure which in addition allows for easier amendment, expansion and upgrading. Such systems are commonly referred to as ‘transfer-based’ systems.

A further difference between systems may be made at this point. It concerns the use of **interlingual** elements. The basic idea is that rather than formulating changes between languages in terms of transfer operations, translation takes place from and into interlingual elements or structures which are in some sense neutral for the languages concerned. At its extreme, the whole system is interlingual, that is all analysis is aimed towards an interlingual representation, and all generation is from interlingual representations. Less extreme is the common use of interlingual elements or features in systems which are otherwise of the 'transfer' type. Whichever approach is adopted however, the translational differences have to be handled and it is these differences I shall be describing.

Methods of analysis and transfer

All translation is a problem-solving activity, choices have to be made continually. The assumption in MT systems, whether fully or partially automatic, is that there are sufficiently large areas of natural language and of translation processes that can be formalised for treatment by computer programs. The basic premise is therefore that the differences between languages can to some extent be regularised. What this means at the practical level is that problems of selection can be resolved by clearly definable procedures. The major task for MT researchers and developers is to determine what information is most effective in particular situations, what kind of information is appropriate in particular circumstances, and whether some data should be given greater weight than others.

In this paper I shall discuss the following types or levels of decision-making:

(a) the use of other specific words in the same phrase or sentence
(b) the use of morphological information
(c) the use of information about syntactic functions and relations
(d) the use of semantic features and relations
(e) the use of knowledge about the subject domain
(f) the use of stylistic preferences

Specific words

Decisions based on specific words are the easiest to apply and are capable of the highest degree of precision. At the same time, however, there is inflexibility since there is no allowance for inflected variation of forms or for the least variation of word order. I shall illustrate with three types of problem: compound nouns, idioms, and metaphors.

All translators are familiar with the need to treat compounds as units to be translated. In many cases an attempt to translate each component of a compound noun would lead to ridiculous results: French *pomme de terre* is not 'apple of earth' but *potato*. If there is a standard equivalent for a particular technical term, then translators are obliged to use it. Many potential problems of homonymy can be averted by the entry of the relevant words in combination with others in dictionaries.

Take the word *light* for example, which can modify another noun in at least three different senses: an adjective 'not heavy', an adjective 'not dark' and a noun 'luminescence or illumination'. In theory, every occurrence could have any one of these senses, but if there are certain words which regularly occur with it, it would seem perverse not to make use of this fact. Thus many MT systems include entries for compounds such as light *ship* and *light bulb*; and indicate directly the target language equivalent (French *ampoule*, German *Glühbirne*). In this way the system can avoid a perhaps lengthy disambiguation process to determine which of the two senses of *bulb* is intended ('plant bulb' or 'pear-shaped glass') and in combination with which of the three senses of light; a process which will have to be done every time the compound is encountered.

For the lay observer, it might seem that the most difficult area for MT must be the apparently unclassifiable variety of idiomatic constructions. It is a view which has support in the apocryphal stories of early MT systems which translated *out of sight out of mind* as 'invisible idiot' and which made such a mess of *The spirit is willing but the flesh is weak*. I shall not rehearse all the variants allegedly produced; the fact that nobody seems to agree of how the phrase was actually translated or indeed which languages were involved should make anyone cautious.

The perceived difficulty of idioms is that the individual words take on meanings and connotations which they do not have in their literal usages. However, it is precisely because most idioms are relatively fixed expressions, consisting of the same words in the same sequence, that they can be easily translated into comparable idioms – or if none exist into a literal equivalent.

Idioms can in fact be treated very much like any compound. For example collocations of *cry* – as we have seen a homonym meaning either 'weep' or 'shout' – can constitute idioms of various lengths and complexities: *cry out, cry off, cry wolf, cry over spilt milk.* Such collocations are sufficient to distinguish the two senses, and at the same time to deal with the idiomatic usages. Hence, however idiosyncratic such idioms as the following (1) may appear it is a relatively simple matter to enter them as units, as fixed phrases in the source dictionary, with their target language equivalents – whether those equivalents are themselves idioms or not.

(1a) bring to heel → mettre au pas
(1b) between a rock and a hard place
(1c) to curry favour → chercher à plaire
(1d) not to see the wood for the trees
(1e) to cry wolf → crier au loup
(1f) to hit the nail on the head → frapper juste
(1g) to oil the wheels → faciliter les choses

The same approach can be taken with many metaphorical usages, e.g. *mouth of river, branch of a bank, flow of ideas, channel of communication, tide of opinion, foot of the mountain, leg of the table*. Like idioms, metaphors of this kind can be treated as fixed compound expressions. We may note that among the European languages there is a common thread of similar formations, so that even if a metaphorical usage is not recorded in the dictionary, it may be possible to produce a 'literal' translation which has the same metaphorical impact. However, it would be a weakness in any MT system if it did not account easily for many metaphors which have become standard expressions of the language.

The advantage of treating certain word combinations as fixed expressions and translating them as units is the considerable saving in processing, particularly the analysis of syntactic structure, and the assurance that the target output will be guaranteed to be correct. There are disadvantages also, however, since idioms can vary in structure, and variation is very common for 'idiomatic' phrasal verbs (2). In other words the identification of idiomatic expressions must often involve morphological and syntactic analysis.

(2a) They had reached the point of no return
(2b) The point of no return had been reached long ago
(2c) They were always crying wolf

The need for syntactic analysis is highlighted by one of the dangers of treating word groups as units. Consider the noun phrase *water pressure*. It would seem reasonable to treat this as a technical term with a fixed translation, i.e. *pression d'eau*.

(3a) The water pressure is low → La pression d'eau est basse
(3b) To lift the well water pressure is obtained from the pump
(3c) Pour faire monter le puits la pression d'eau est obtenue à l'aide de la pompe.
(3d) Pour faire monter l'eau du puits la pression est obtenue à l'aide de la pompe.

While there would clearly be no problem in a sentence such as (3a), there would be unfortunate consequences in a sentence such as (3b). Here the two words *water* and *pressure* occur in different clauses. The desired translation is (3c) and not (3d). What is needed is some kind of syntactic analysis which identifies phrase boundaries before attempts are made to locate compounds in the dictionary. It is reasonable to suppose that if a MT system produces mistakes of this nature, then syntactic analysis is being treated as subsidiary to compound identification, and not (as it probably should be) vice versa.

A relatively minor complication is that occasionally idioms and metaphorical expression can be interpreted literally, e.g. *to oil the wheels* could be metaphorically *faciliter les choses* or literally graisser *les roues*. If there is only one translation given in the dictionary then the literal interpretation will be missed. It is not a problem to be exaggerated, since it is an easy matter for translators to make the appropriate revision. Indeed this comment applies to nearly all translations of fixed expressions. They are both easy for the MT system and easy for the reviser to change.

Morphological analysis

It is a truism to say that one of the most straightforward operations of any MT system should be the identification and generation of morphological variants of nouns and verbs. There are basically two types of morphology in question: inflectional morphology, as illustrated by the familiar verb and noun paradigms (French *marcher, marche, marchons, marchait, est marché*, etc.), and derivational morphology, which is concerned with the formation of nouns from verb bases, verbs from noun forms, adjectives from nouns, and so forth, e.g. *nation, nationalism, nationalise, nationalisation*, and equivalents in other languages.

It should be stressed that any MT system should as a minimum be capable of recognising morphological forms and of generating them correctly. However, the alignment of equivalences between the verb forms between languages is another matter, particularly when modal forms are involved (*must, might, devoir, falloir, mögen, dürfen*, etc.)

In general, a MT system which cannot go beyond morphological analysis will produce little more than word for word translations. It may cope well with compounds and other fixed expressions, it may deal adequately with noun and verb forms in certain cases, but the omission of any treatment of word order will give poor results. Nevertheless, the output of such programs can serve useful purposes in certain circumstances. A specialist in a particular subject area who knows something of the grammatical characteristics of the source language could well find that comprehension of the gist of the text was not impossible. There is in fact evidence that scientists have found the crude output of some of the early Russian-English systems to be of considerable assistance in keeping up with research developments. Another use for such crude output can be as ‘pre-translation’ versions for experienced translators. They have the assurance that the technical vocabulary has been correctly translated, and they retain sufficient indication of the original sentence structure to permit reworking for good quality translation.

Syntactic structures

I shall not attempt to describe the various approaches to syntactic analysis, and the complexities of structural transfer and generation. What I shall do is simply point out some of the major areas of difficulty, and the chief reasons for those difficulties.

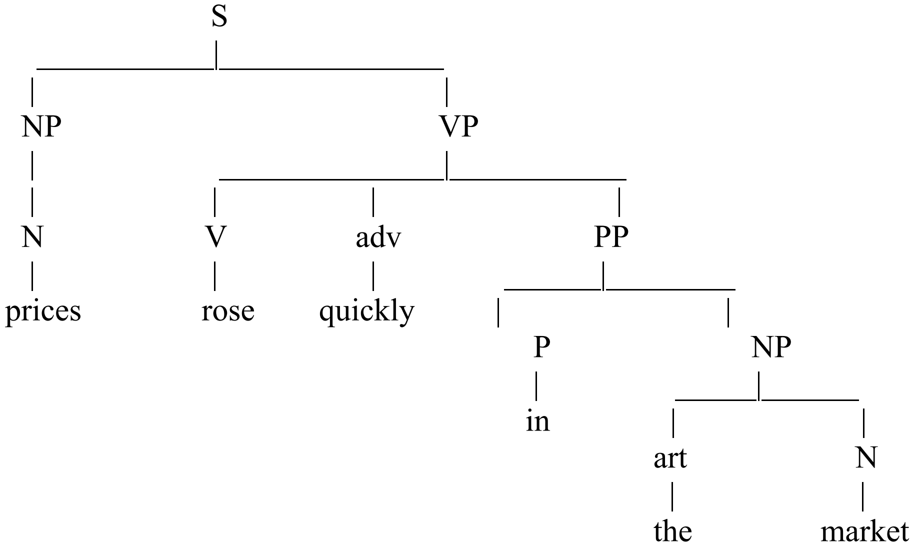
The basic structural features are those of dependency and constituency. Examples of dependency are the relations between adjectives and the nouns they modify and between subject nouns and the main verbs of clauses. Any MT system should be able to identify such relations in languages such as French and German on the basis of gender agreement: *les jeunes filles sont venues, die meisten Frauen sind nicht gekommen*. There are of course some complexities which are not easy and we shall come to them in a moment. In English the lack of overt markers of dependency or the ambiguity of those markers which do exist means that greater weight has to be given to the identification of constituency groups, e.g. noun phrases, verb phrases, prepositional clauses and phrases, etc.

Syntactic analysis is based largely on the identification of grammatical categories: nouns, verbs, adjectives. For English, the major problem is the categorial ambiguity of so many words, as already illustrated with the word *light*. In essence, the solution is to look for words which are unambiguous as to category and to test all possible syntactic structures. In the case of a sentence such as:

(4) Prices rose quickly in the market

Each of the words *prices*, *rose*, and *market* can be either nouns or verbs; however, quickly is unambiguously an adverb and the unambiguously a definite article, and these facts ensure the unambiguous analysis as a phrase structure (5), where *prices* is identified as a subject noun phrase, *in the market* as a prepositional phrase, and rose *quickly* as part of a verb phrase. (Note that this particular analysis is not one necessarily found in any MT system and would not be adopted by many syntax theories.)

(5)



The example demonstrates that without using any semantic information it is in fact possible for homonyms to be disambiguated by syntactic analysis. As another example, the word *return* can mean either ‘go back’ or ‘give back’ (with correspondingly different translations in other languages: *rendre*, *retourner*, zurückgehen, *zurückgeben*). The two meanings can often be distinguished by the presence or absence of a direct object (6):

(6a) She returned to the office → Sie ging in das Büro zurück
(6b) She returned the book → Sie gab das Buch zurück

The identification of structural context can also be sufficient for translation of words which are not homonymic in the source language but which have more than one possible equivalent in the target language. A familiar example is the verb *know* (7). Although there are other factors involved, the identification of the structures in which *know* occurs can assist in the choice between savoir and *connaître* and between *wissen* and kennen, as shown in (7a) and (7b).

(7a) I know his brother → je connais son frère; ich kenne seinen Bruder
(7b) I know what he is called → je sais ce qu’il s’appelle; ich weiß wie er heißt

Structural changes are so common when translating from one language into another that the 'low-level' ordering of basic elements – nouns, verbs, object nouns, adverbs, adjectives – should be expected from any MT system. In French output, for example, the pre-nominal pronouns (*me, vous, le, lui, leur*, etc.) should be in the correct order and discontinuous *ne...plus*, ne*...rien*, etc. should be correctly placed. In German, the system should generate the correct placement of verbal elements in main clauses (*er hat es gestern empfangen*) and in subordinate clauses *(...daß er es gestern nicht empfangen hatte*), and so forth. This must be regarded as the minimum requirement, and any MT system which fails in this respect must be suspected of deficiencies elsewhere, probably of a graver nature.

Many structural changes are linked with specific lexical items, such as the *know* examples above and the following examples with *like* (in various meanings):

(8a) Young people like this music → Cette musique plaît aux jeunes gens
(8b) The boy likes to play tennis → Der Junge spielt Tennis gern
(8c) The boy was like his sister → Le garçon resemblait à sa soeur
→ Der Junge ähnelte seine Schwester

More complex restructuring may well be beyond the capacity of many cheaper systems, such as the treatment of complex German pre-nominal modifying phrases and clauses (9). For translation into English post-nominal structures, (9c) or (9d), the system must be capable in effect of recognising the equivalence of the structure (9a) with a full relative clause (9b), and when generating the English structure (9d) it must 'know' what parts of relative clauses (9c) can be omitted without loss.

(9a) Die in den letzten Jahren entwickelten Technologien....
(9b) Die Technologien, die in den letzten Jahren entwickelt geworden sind,...
(9c) The technologies, which have been developed in recent years,...
(9d) The technologies developed in recent years...

The example demonstrates the need for syntactic analysis and transfer to operate at a 'deeper' level that surface relationships (such as those in (5) above.) What is often meant by 'deeper' analysis is the extraction of implicit functional relationships. In the following examples (10) it is clear that the subject of the subordinate verb (*visit*) is *Mary* in the first sentence (10a) while it is *John* in the second (10b)

(10a) John persuaded Mary to visit his father
(10b) John promised Mary to visit his father

Semantic roles and features

The recognition of implicit relations may well require access to semantic information. It is common to identify two types: semantic roles and semantic features. By the semantic roles in a structure is meant the specific relationships of nominal elements (entities) to verbal elements (actions or states): a particular noun may be the 'agent' of an action, another may be the 'instrument' (or means), another may be the 'recipient', and another may refer to the 'location', and so forth. Such roles can apply to both full sentences or to phrases (11):

```
(11a) In Europe the rivers are polluted by chemicals from industry
          loc            rec     [action]         agent             source
(11b) the pollution of rivers by chemicals in Europe
          [action]          rec     agent          loc
```

(11c) the industrial pollution of European rivers
source [action] loc rec

Unfortunately, there is no universally agreed set of semantic roles which can be applied without difficulty to any language. Developers of MT systems are usually obliged to draw up their own list. However, the principal difficulty is the identification of roles. In English the main indicators are the prepositions, but these can be ambiguous as to the role expressed; *with* can indicate instrument, manner or context:

(12a) The bottle was opened with a corkscrew
(12b) The bottle was opened with difficulty
(12c) The bottle was opened with the meal

In many languages roles are often expressed by case endings (e.g. Russian and German) - hence they are also called ‘deep cases’. But these too can be ambiguous: German *-en* can indicate a dative singular noun, any of the plural noun cases and a large number of adjectival endings.

Despite the difficulties, analyses of case relations are almost essential when translating between languages of widely diverging structures, such as Japanese and English; compare, for example (13a) and (13b):

(13a) The earthquake destroyed the buildings
inst [action] obj
(13b) jishin de kenbutsu ga kowareta
earthquake by-means-of buildings (subj) collapsed

The instrumental role of the subject noun *earthquake* in English (12a) must be recognised if the appropriate Japanese structure is to be generated with *de* (by-means-of) following *jishin* and ga (subject marker) following *kenbutsu*. Similarly, the treatment of the *like* example in (8a) above can be assisted by aligning the semantic roles of the languages, i.e. both *young people* and *jeunes gens* are ‘recipients’ and both *this music* and cette *musique* are ‘sources’.

Semantic features refer to labels such as ‘human’, ‘animate’, ‘liquid’, ‘young’, etc. assigned to lexical elements. They can be used either in conjunction with semantic roles or independently. For example, for the translation of English *eat* into German it might be considered useful to distinguish between ‘human’ agents (14a) and ‘non-human’ (14b)

(14a) The boy ate the banana → Der Junge hat die Banane gegessen
(14b) The monkey ate the banana → Der Affe hat die Banane gefressen

Such features have to be assigned to all relevant nouns (i.e. all that could be subjects of the verb eat); and can be used in other sentences where choices between human and non-human have to be made. As with semantic roles, however, there is no established set of features which can be applied to every language; MT developers have complied their own lists, some are minimal and rigidly controlled, others are extensive or not applied consistently. There are obvious dangers with inconsistent application, which can be increased if individual users are able to add vocabulary items with semantic features (as they can in a number of commercially available systems). In addition, there is the risk of over-specification. To take a somewhat trivial example, if it were held that the subjects of the verbs *drink* and *die* could only be ‘animate’, then the system would reject the following sentences: *The car drinks petrol*, and *The secret died with him*.

The application of semantic roles and features in the analysis of noun compounds is no easy matter. As is well known, in English nouns can be modified by other nouns functioning as quasi-adjectives. The lack of explicit markers of the relationships can cause problems, because in many other languages these relationships have to be expressed – e.g. by case endings, or by prepositions. For example, in an English sequence adjective-noun-noun the adjective can modify either the first noun (15a) or the second noun (15b).

(15a)
private company legislation → legislation for private company
(adj n) n
hydraulic brake fluid → fluid for hydraulic brake(s)
(adj n) n
(15b)
private water company → water company which is private
adj (n n)
diluted brake fluid → brake fluid which is diluted
adj (n n)

In theory, each of these groupings is possible for every such adjective-noun-noun sequence since syntactic analysis alone cannot select the intended modification. But the application of semantic features and roles would require the detailed specification of what types of adjectives and nouns can modify which particular types of nouns and noun phrases (whether derived from verb forms or not). Not surprisingly, MT systems adopt the easiest solutions, namely either to leave the decision to a human assistant during translation or to a human reviser after translation, or to include noun compounds in the dictionary, i.e. water *company* and *hydraulic brake* are treated as 'fixed' expressions (with the potential consequences of incorrect analysis already mentioned.)

There is a further possibility and this is to make reference to a knowledge bank containing data about the subject domain. In the next section I shall say something about this last option.

Real world knowledge

While semantic features and roles combined with syntactic information can go a long way in resolving ambiguities in the source language and in deciding among translation variants, there are numerous instances where what is apparently needed is knowledge about the things and events being referred to. Take some simple problems of coordination:

(16a) old men and women → les vieux et les vieilles
or: les vieux et les femmes
(16b) pregnant women and children → des femmes enceintes et des enfants
not: des femmes et des enfants enceintes
(16c) smog and pollution control
(16d) Smog and pollution control are important factors
Smog and pollution control is under consideration
(16e) The authorities encouraged smog and pollution control.

In (16a) we have no idea, out of context, whether old applies to both men and women or only to men, and either translation into French is possible. But in (16b) we do know that pregnant cannot apply to children; it is part of our knowledge about women. This knowledge needs to be incorporated in the MT dictionary in some way, probably by limiting the use of pregnant to nouns with the semantic features 'female' and 'mature'. In (16c) there are also two possible interpretations, which could be disambiguated with morphological and syntactic information (16d).

However, with (16d) we need information about what is ‘reasonable’ behaviour by those in authority; we do not these days expect them to encourage industry to cause smog.

Similar problems arise with relative clauses:

(17a) Peter mentioned the book I sent to Mary
→ mentioned the book (which I sent to Mary)
→ mentioned (to Mary) the book (which I sent)
(17b) We will meet the man you told us about yesterday
(17c) We will meet the man you told us about tomorrow

The first sentence (17a) is ambiguous: either the book itself was sent to Mary or the sending of a book to someone else was mentioned to Mary. It is an ambiguity which cannot be solved out of context. In (17b) and (17c), by contrast, there is information which enables the correct analyses to be made: in (17b) the word *yesterday* is attached to the clause with the past tense *told*, and in (17c) the word *tomorrow* is attached to the clause with the future tense verb *will meet*. This can be regarded as semantic information, i.e. matching features of ‘past’ and ‘future’, or as information about the phenomenon of time.

More complex yet are examples such as the following, which have implications for the correct generation of French and German translations.

(18a) Having solved the problem, he went to bed
(18b) Having forgotten his book, he went back to fetch it

The relationship between the two clauses is implicit, and differs in the examples: in (18a) it is a temporal relationship, in (18b) it is a causal relationship. Since French and German do not permit constructions of this kind, the relationships must be made explicit:

(19a) After he had solved...
→ Nachdem er das Problem gelöst hat, ...
→ Après qu'il a resolu le problème, ...
(19b) As he had forgotten...
→ Da er das Buch vergessen hat,...
→ Puisqu'il a oubli le livre,...

Probably all MT systems have difficulties with this kind of construction. An examination of the semantic features of the verbs may suffice on occasions, but in many cases it will not. What seems to be involved is knowledge about human behaviour, the system needs to have some kind of human-like ‘understanding’.

We are led therefore to the argument that good quality translation is not possible without understanding the reality behind what is being expressed, i.e. translation goes beyond the familiar linguistic information: morphology, syntax and semantics. The need is particularly striking in the treatment of pronouns. Human translators have virtually no problems with pronouns, and it must seem strange to many that while MT systems can deal quite well with complex idioms and certain complex structures, they all seem to have great difficulties with pronouns. Why do we get such errors as *die Europäische Gemeinschaft und ihre Mitglieder*, rendered as *the European Community and her members*? The problem is that the antecedent of pronouns must be identified; the default translation of *ihr* as *her* does not work. The antecedent is often the immediately preceding noun, but certainly not always. Compare the following:

(20a) The monkey ate the banana because it was hungry

Der Affe... er...
(20b) The monkey ate the banana because it was ripe
... die Banane... sie...
(20c) The monkey ate the banana because it was tea-time
es...

Semantic information may well suffice: in (20a) the association of hunger with an ‘animate’ being should link it to the monkey and produce the correct *er* in German; in (20b) the association of *ripe* with a ‘fruit’ should enable the correct generation of sie; but in (20c), the pronoun refers not to a specific element but to the whole event, and has to be rendered by the impersonal es. Here non-linguistic information is called upon. As a more extreme example consider:

(21a) The soldiers shot at the women and some of them fell
(21b) The soldiers shot at the women and some of them missed

We can identify the antecedents of *them* from our knowledge of what happens in the ‘real world’. It is knowledge which is very difficult to formulate in computer programs, but it is essential if the correct translations are to be made:

(22a) Les soldats ont tiré sur les femmes et quelques-unes sont tombées
(22b) Les soldats ont tiré sur les femmes et quelques-uns ont râté

The clear implication is that what is required for the translation of the more intractable problems of analysis and transfer is the availability of a knowledge bank of information which can be referred to during the translation process. It is the approach commonly referred to as that of Artificial Intelligence (AI). For example, given a sentence such the following occurring in documents relating to computer hardware.

(23) Remove the tape from the disk drive

The word *tape* can potentially refer to a ‘magnetic tape’ or an ‘adhesive tape’. An AI-based system would check in its knowledge bank which is most plausible in this context, i.e. it would seek to answer the question whether magnetic tapes can be removed from disk drives, or whether disk drives can contain or have as parts items which are magnetic tapes. If not, then it may check whether ‘adhesive tape’ is plausible, i.e. whether disk drives are things which can be packaged using this item. Clearly, the knowledge bank must contain highly structured information about a wide range of real phenomena, even when documents deal with a quite narrow domain.

As far as MT is concerned, this approach is to be found in only a few experimental systems. There are some vendors of MT systems who claim that they incorporate AI techniques, but what this means generally is that they have made extensive use of semantic features and roles. There is of course a close link between semantic features and the 'real' attributes of objects, but AI goes beyond the common use of semantic features; it involves the formalisation and identification of expected behaviour in domain-specific situations, the inferring of unstated assumptions, and the specification of common sense as well as subject knowledge.

The principal reasons for the absence of knowledge banks in MT systems are probably obvious enough. Coverage of any documents other than those within a narrow subject range would clearly require databases of massive proportions. While the computer hardware and the computer software for fast access may well both be already available, the databases are not. These would demand many years of difficult and complex work by many researchers. Therefore, it is not surprising that MT systems are based on well known techniques of syntactic and semantic analysis and transfer.

Stylistic matters

One of the most distinctive features of texts produced by MT systems is their unnatural literalness. In general, they adhere too closely to the structures of source texts. Of course, human translators can be guilty of this fault as well – although Newmark (1991) considers literalness to be desirable in literary and authoritative texts, as long as the result is in the appropriate style. However, the aim in technical translation is generally to produce texts which read as if they were originally written in the target language. It is quite evident that MT systems do not achieve this goal. Indeed, it can be argued that they should not aim for idiomaticity of this order, if only because recipients of MT output may be led to assume complete accuracy and fidelity in the translation. It does not need stressing that readability and fidelity do not go hand in hand: a readable translation may be inaccurate, and a faithful translation may be difficult to read.

Nevertheless, it is reasonable to aspire to some degree of idiomaticity in MT output. It would certainly help MT to become more acceptable. What we are talking about in this respect is primarily procedures for the generation of text in the target language.

An area of particular difficulty for most MT systems is the generation of prepositions. Consider the following:

(24a)

es hängt von dem Direktor ab	it depends on the director
er glaubt nicht an Gott	he does not believe in God
er hat sich an das Klima gewöhnt	he got used to the climate

(24b)

il est certain de réussir	he is certain to succeed
il est capable de résister	he is capable of resisting
il vient de Paris	he comes from Paris
il partit de nuit	he left at night
il partit de bonne heure	he left in good time
le meilleur élève de la classe	best pupil in the class

Some of the differences can be handled at the transfer stage. The German prepositions (24a) are bound closely to particular verb forms, and it is clearly possible to transfer the verb and its appropriate preposition as a unit into an English verb and a preposition (*glauben an → believe in*). When we look at the French examples (24b) some could well be dealt with in a similar manner but often it seems that it is idiosyncrasies of English which are more significant (*at night, in good time*)

A similar situation arises with certain lexical transfers. Consider some of the possible translations of French grand and German *gross* into English.

(25)

grand succès/grosser Erfolg	big/great success
grosses Tier	big/large animal
grosse Fläche	large/great expanse
grande vitesse/grosse Geschwindigkeit	high speed
grande question/grosse Frage	big/great question
grosse Hände	big/large hands
grand chef/grosser Führer	great leader
grand espoir/grosse Hoffnung	great hope

It is difficult to suppose that French and German speakers consider *grand* or *gross* to be polysemous (although it might just be possible that they recognise a distinction between its literal use, referring to size, and its metaphorical extension, referring to achievement.) It is more plausible to see it as a matter of stylistic variety to be handled in the generation of English. Even more so in the case of French *rapide* which can be fast*, quick, rapid, swift*, etc. according to the specific English noun concerned (*rapid development, swift progress, fast memory, quick response, swift action, fast access*, etc.). If there are any semantic differences among the English adjectives they are elusive.

In recent years MT researchers have paid increasing attention to the methods for producing idiomatic output. A popular approach at present is the introduction of various kinds of statistical or probabilistic weighting of target language structures and lexical items. In the case of *grand* and *rapide*, for example, a database of noun phrases in English and French translations (e.g. a greatly expanded list of examples as in (25) above) would enable appropriate renditions to be either extracted directly or derived by probabilistic pattern matching procedures. This approach has been made possible by the growing numbers of large textual databases, of all kinds of documents, and by improvements in computer analysis and processing of large databases.

Until such research advances much further however, current MT systems are unlikely to do more, as far as the production of idiomatic output is concerned, than certain minimal kinds of transposition. In English, indirect objects can occur before direct objects (26a), particular if they are pronoun (26b), but not if they are longer or more complex (26c). In the latter case, only the inverted form is acceptable (26d).

(26a) He gave the assistant the money
(26b) He gave her the money
(26c) He gave the assistant in the flower shop on the corner of the street the money.
(26d) He gave the money to the assistant in the flower shop on the corner of the street

Stylistic preferences of this nature are easily handled. Most are not. For example, there is a tendency in many English document types towards a nominal style which is not shared in other languages. Whereas an English author might write (27a), the preference in many other languages would be something more like (27b).

(27a) The possibility of rectification of the fault by the insertion of a valve is discussed.
(27b) We discuss whether it is possible to rectify the fault by inserting a valve.

Changes to take account of such stylistic preferences are likely to remain beyond the capacity of most MT systems for the foreseeable future.

Conclusions

This account has, of course, by no means exhausted all the areas in which MT systems may have difficulties. Some of these are nothing to do in essence with the problems of translation as such. Obviously, gaps in dictionaries are particularly important, but simple mistakes of spelling and of grammar can lead to systems failing completely to produce any version or to producing something which is incomprehensible. However, even here researchers are developing systems which can cope with common spelling and grammar mistakes (e.g. *supercede*, procedings, incidently*, none of them were present*, *he did not here them*, *you must try and see the exhibition*, etc.)

Since the major problems of MT systems concern ambiguity, homonymy and alternative structures, it has long been recognised that one of the best ways of ensuring good MT output is to limit the amount of choice in the actual texts submitted to the system or to limit the system itself to specific text types or subject areas. The latter is exemplified by the well known Meteo system, which was designed for meteorological texts and for nothing else (Chandioux 1989). The former is being adopted by an increasing number of MT users, who require texts to conform to certain restrictions of vocabulary and syntax: certain words are to be used in one meaning only, and complex structures are to be avoided. For example, the word *replace* can mean either 'put back' or 'exchange' (28); to avoid the ambiguity, it may be restricted to only the first sense and *substitute* used instead for the second meaning.

(28a) Remove part A and replace it after cleaning
(28b) Remove part A and replace with part B.

Complex sentences may also be avoided, in order to ease problems of MT analysis. Thus (29a) may be rephrased as (29b):

(29a) Loosen the main motor and drive shaft and slide back until touching back plate
(29b) Loosen the main motor. Loosen the drive shaft. Slide both parts until they touch the back plate.

I shall say no more here about this approach, since it is clearly a way of avoiding problems rather than tackling them (e.g. Pym 1990), and my topic in this paper has been the difficulties encountered by MT systems when they do tackle them.

There are well-tested and familiar methods for word recognition, for morphological segmentation and for syntactic analysis. The use of semantic features and roles is also well researched and reliable. With these techniques it is possible to deal with a wide range of linguistic phenomena with reasonable success – but not always without problems. As I have briefly illustrated, among phenomena which can be relatively easily handled are: idioms and fixed expressions, phrasal verbs, basic word order (both in analysis and in generation), metaphors (when identifiable by specific words), the morphological and the syntactic disambiguation of homonyms, and the resolution of ambiguities by the use of simple semantic features.

There remain, however, many phenomena of greater difficulty. Some may not occur often in certain text types and some may not be critical for certain users (i.e. they can be handled easily in post-editing or in interactive modes of operation) – how much difficulty they cause depends largely on local circumstances. Among these relatively more difficult phenomena are: prepositions, tense and modality, coordination, subordinate clauses, pronouns, complex sentences, and stylistic variants (both lexical and structural). Various methods and techniques are currently being researched which may provide solutions; here I have mentioned knowledge-based approaches (Nirenburg et al. 1992), statistical methods (Brown et al. 1990), and the use of text corpora of example translations (Sadler 1989). In addition, there are new approaches to linguistic formalisms of many kinds, and we should not omit to mention the increasing attention paid to the practical experiences of professional translators and to developments in translation research.

I have attempted in this paper to outline the nature of the problems faced by designers MT systems, and why these problems are relatively easy or difficult to tackle. Some difficult problems may prove to be inherently unsolvable. Some are certainly intractable with present methods and at the present stage of knowledge. For others there are good prospects of viable approaches; research continues and we can hope for gradual if not dramatic improvements in the future.

References

Brown, P. et al. (1990): A statistical approach to machine translation. *Computational Linguistics* 16 (1990), 79-85.

Chandioux, J. (1989): Météo, 100 million words later. In: Hammond, D.L. (ed.) *American Translators Association Conference 1989: Coming of age*. Medford, N.J.: Learned Information, 1989, pp. 449-453.

Hutchins, W.J. & Somers, H.L. (1992): *An introduction to machine translation*. London: Academic Press, 1992.

Newmark, P. (1991): *About translation*. Clevedon: Multilingual Matters, 1991.

Nirenburg, S. et al. (1992): *Machine translation: a knowledge-based approach*. San Mateo, Ca.: Morgan Kauffmann, 1992.

Pym, P. (1990): Pre-editing and the use of simplified writing for MT. In: Mayorcas, P. (ed.) *Translating and the computer 10*. London: Aslib, 1990.

Sadler, V. (1989): *Working with analogical semantics*. Dordrecht: Foris, 1989.