[from: *Sprog og Erhverv*, no.3, 1992, pp.4-8]

TRANSLATION: WHAT COMPUTERS CAN DO

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

With the well known demand for translation of texts of all kinds growing at an ever faster rate, it is clear that few translators in the future will be able to survive without some computer-based assistance. The most common form of this assistance will be in the form of machine aids: automatic dictionaries, optical character readers, multilingual publishing faciltities, and the like. However, the possibilities of automatic translation should not be overlooked. It is unlikely that there will be in the near future any *machine translation* (MT) system capable of producing good quality output without human intervention. Translators need have no fear that machines will take over their jobs, but they need to be aware of what computers are already capable of and how they can contribute to the translation effort - whether with or without the assistance of translators. In this paper I shall briefly describe the basic design features of machine translation systems and how they tackle certain linguistic problems.

Although the ideal aim of MT systems might be to produce translations as good as those from the best human translators, in practice the output has to be revised (or *post-edited*) for most recipients. In this respect MT does not differ from the output of most human translators which is normally revised by a second translator before dissemination. However, the types of errors produced by MT systems do differ from those of human translators (incorrect prepositions, articles, pronouns, verb tenses, etc.). Post-editing is the norm, but in certain circumstances MT output may be unedited or only lightly revised, e.g. if it is intended only for specialists familiar with the subject field of the text. Unrevised output might also serve as a rough draft for a human translator, often referred to as a 'pre-translation'.

The translation quality of MT systems may be improved by imposing certain restrictions on the input. The system may be designed, for example, to deal with texts limited to the *sublanguage* (vocabulary and grammar) of a particular subject field (e.g. biochemistry) and/or document type (e.g. patents). Alternatively, input texts may be written in a *controlled language*, which restricts the range of vocabulary, avoids homonymy and polysemy, and eliminates complex sentence structures. A third option is to mark input texts (*pre-edit*) to indicate prefixes, suffixes, word divisions, phrase and clause boundaries, or to differentiate grammatical categories (e.g. to distinguish the noun *convict* from its homonymous verb *convict*.) Finally, the system itself may refer problems of ambiguity and selection to human operators (usually translators) for resolution during the processes of translation itself, i.e. in an *interactive* mode.

In overall system design, there have been three basic types of systems. The first (and historically oldest) type is generally referred to as the *'direct translation'* approach: the MT system is designed in all details specifically for one particular pair of languages, e.g. Russian as the language of the original texts, the *source language* (SL), and English as the language of the translated texts, the *target language* (TL). Translation is direct from the SL to the TL text; the basic assumption is that the vocabulary and syntax of SL texts need not be analysed any more than strictly necessary for the resolution of ambiguities, the correct identification of TL

expressions and the specification of TL word order; in other words, SL analysis is oriented specifically to one particular TL.

The second basic design strategy is the *'interlingua'* approach, which assumes that it is possible to convert SL texts into representations common to more than one language. From such interlingual representations texts are generated into other languages. Translation is thus in two stages: from SL to the interlingua (IL) and from the IL to the TL. Procedures for SL analysis are intended to be SL-specific and not oriented to any particular TL; likewise programs for TL synthesis are TL-specific and not designed for input from particular SLs. Interlinguas may be based on a 'logical' artificial language, on an auxiliary language such as Esperanto, on a set of semantic primitives common to all languages, or on a 'universal' vocabulary.

The third basic strategy is the less ambitious *'transfer'* approach. Rather than operating in two stages through a single interlingual representation, there are three stages involving underlying (abstract) representations for both SL and TL texts. The first stage converts SL texts into abstract SL-oriented representations; the second stage converts these into equivalent TL-oriented representations; and the third generates the final TL texts. Whereas the interlingua approach necessarily requires complete resolution of all ambiguities in the SL text so that translation into any other language is possible, in the transfer approach only those ambiguities inherent in the language in question are tackled; problems of lexical differences between languages are dealt with in the second stage (transfer proper).

The main linguistic problems encountered in MT systems may be treated under four main headings: lexical, structural, contextual, and pragmatic or situational. In each case the problems are primarily caused by the inherent ambiguities of natural languages and by the lack of direct equivalences of vocabulary and structure between one language and another. Many examples could be given, some English ones are: homonyms (*cry* as 'weep' or 'shout', *bank* as 'edge of river' or 'financial institution') require different translations (e.g. in German *weinen*: *rufen*; *Ufer*: *Bank*); nouns can function as verbs (*control*, *plant*, *face*) and are hence 'ambiguous', since the TL may well have different forms (e.g. French *contrôle*: *diriger*, *plant*e: *planter*, *face*: *affronter*). A polysemous word such as English *field* has often many possible translations, e.g. in Japanese: *hatake* (field for crops), *nohara* (open space), *kyougiba* (sports field), *bun'ya* (sphere of activity), etc. In many cases, target languages make distinctions which are quite absent in the source, e.g. English *river* can be French *fleuve* or *rivière*, German *Fluss* or *Strom*.

In many cases, differences between SL and TL vocabulary are also accompanied by structural differences. A familiar example is the translation of the English verb *know* into French or German, where there are two verbs which express 'knowledge of a fact' (*connaître* and *kennen*) and 'knowledge of how to do something' (*savoir* and *wissen*):

(1) I know the man - Je connais l'homme; Ich kenne den Mann.

(2) I know what he is called - Je sais ce qu'il s'appelle;
Ich weiss wie er heisst.

Translation into unrelated languages typically involves considerable structural change. For example, the English sentence

(3) must be completely reformulated in Japanese (4):

(3) The earthquake destroyed the buildings

(4) Jishin de kenbutsu ga kowareta
Earthquake-by buildings collapsed
i.e. 'The buildings collapsed due to the earthquake'

Various aspects of syntactic relations can be analysed. There is the need (a) to identify valid sequences of grammatical categories, (b) to identify functional relations: subjects and objects of verbs, dependencies of adjectives on 'head' nouns, etc., (c) to identify the constituents of sentences: noun phrases, verb groups, prepositional phrases, subordinate clauses, etc. Each aspect has given rise to different types of parsers, each with their strengths and weaknesses. Modern MT

systems often adopt an eclectic mixture of parsing techniques, now often within the framework of a 'unification grammar' formalism.

The most serious weakness of all syntactic parsers is precisely their limitation to structural features. An English prepositional phrase can in theory modify any preceding noun in the sentence as well as a preceding verb:

(5) The car was driven by the teacher with great skill

(6) The car was driven by the teacher with defective tyres

(7) The car was driven by the teacher with red hair

In (5) the phrase *with great skill* modifies the verb phrase *was driven*; in (6) *with defective tyres* is attached to *the car*; and in (7) *with red hair* is an attribute of *the teacher*. However, these attachments are based on semantic or pragmatic information (e.g. knowledge that cars do not have red hair). But syntactic analysis can go no further than offer each possibility, and the specific relationship has to be identified by later semantic analysis involving lexical and situational context.

To overcome some of these problems, many parsers now include the identification of case relations. Consider, for example:

(8) The house was built by a doctor for his son last year.

In this sentence, the Agent of the action ('building') is *a doctor*, the Object of the action is *the house* the Recipient (or Beneficiary) is *his son* and the Time of the action is *last year*. Many languages express these relations explicitly: suffixes of Latin, German, Russian nouns *(-ibus*, *-en*, *-ami*), prepositions of English and French (*to*, *à*), particles of Japanese (*ga*, *wa*); but they are often implicit (as in English direct objects). There are rarely any direct correspondences between languages and most markers of cases are multiply ambiguous in all languages, cf. *with* expressing Manner (5) or Attribute (6 and 7). Nevertheless, there is a sufficient regularity and universality in such 'case relations' to have encouraged their widespread adoption in many MT systems.

There is also some agreement about the use of semantic features, i.e. the attachment of such categories as 'human', 'animate', 'liquid' to lexical items and their application in the resolution of ambiguities. For example, in:

(9) He was beaten with a club

In this sentence, the 'social' sense (meeting place) of *club* is excluded by the verb-type which requires an 'inanimate' Instrument, i.e. it must have the 'weapon' meaning.

Few operational MT systems involve any deeper levels of semantic or pragmatic analysis, yet the resolution of many linguistic problems clearly transcends sentence boundaries. A common and persistently difficult problem involves the use of pronouns. Consider the following:

(10) The soldiers shot the women. They were buried next day.

We know that the pronoun *they* does not refer to *soldiers* and must refer to *women* because we know that 'shooting' implies 'killing' and 'injury' or 'death' and that 'death' is followed (normally) by 'burial'. This identification is crucial when translating into languages where the pronoun must indicate whether the referents are male or female (e.g. French *elles* and *ils*.) Such examples demonstrate that the disambiguation and correct selection of TL equivalents would often seem to be impossible without reference to knowledge of the (non-linguistic) features and properties of the actual objects and events described. Recent advances in Artificial Intelligence have encouraged the investigation of knowledge-based MT systems, at least for systems restricted to specific domains. For example, in a system designed for translating texts in computer science and data processing the English word *tape* could mean either magnetic tape or adhesive tape. In the following sentence (11), reference to the knowledge base should establish that only the 'adhesive tape' interpretation is possible since diskettes do not contain magnetic tapes which can be removed.

(11) Remove the tape from the diskette

In view of the linguistic limitations of MT systems it should be clear that the most suitable texts are either those of a technical or scientific nature, where there is often a high degree of direct terminological equivalence and where problems of homonymy and polysemy can be reduced by the restriction of dictionaries to specific subject domains, or administrative texts with a high degree of repetition, where stylistic considerations are unimportant (e.g. the minutes of meetings, internal reports, etc.) Obviously unsuitable are literary and philosophical texts, where nuances of vocabulary and cultural and stylistic factors play an important role; and equally unsuitable are texts with particularly complex sentence structures, e.g. patents and legal documents. The suitability of texts intended for publication depends on various economic factors, such as whether input texts are in machine-readable form, whether long documents change little between editions (e.g. operational manuals for equipment), whether a great deal of terminology work has to be done, and so forth.

The immediate future is likely to see progessive improvements in both batch and interactive systems, both for translators and for non-translators. In particular, there are likely to be systems for monolinguals who wish to communicate simple business messages in languages they do not know. There will remain, however, a substantial demand for translation which automated systems will not be able to satisfy. (For instance, MT vendors have naturally concentrated on systems for the major commercial languages, English, French, Japanese, Spanish; languages such as Danish have been largely ignored.) It is essential for translators that they become more closely involved, not only so that they know what facilities have been developed and how they may be used cost-effectively in their own situations but also so that they know what MT research is going on and how they may influence the direction it is taking. The recently established International Association for Machine Translation will provide a forum for the exchange of information and views among all with an interest in the MT field - researchers, developers, manufacturers, vendors, translators, users, etc. - and you are encouraged to join its regional association, the European Association for Machine Translation, by contacting Tamara Wehrli (Secretariat EAMT, ISSCO, 54 route des Acacias, CH-1227 Carouge, Geneva, Switzerland).