[from: *Encyclopedia of Computer Science*, 3rd edition, editors: Anthony Ralston, Edwin D.Reilly
(New York: Van Nostrand Reinhold, 1993), pp.733-738]

LANGUAGE TRANSLATION

John Hutchins

History of MT

Within a few years of the first appearance of the 'electronic calculators' research had begun on using computers as aids for translating natural languages. The major stimulus was a memorandum in July 1949 by Warren Weaver, who after mentioning tentative efforts in Great Britain (by Booth and Richens) and in the United States (by Huskey and others in Los Angeles) put forward possible lines of research. His optimism stemmed from the war-time success in code-breaking, from developments by Shannon in information theory and from speculations about universal principles underlying natural languages, "the common base of human communication". Within a few years research had begun at many US universities, and in 1954 the first public demonstration of the feasibility of machine translation (MT) was given (a collaboration of IBM and Georgetown University). Although using a very restricted vocabulary and grammar it was sufficiently impressive to stimulate massive funding of MT in the United States and to inspire the establishment of MT projects throughout the world.

The earliest systems consisted primarily of large bilingual dictionaries where entries for words of the source language (SL) gave one or more equivalents in the target language (TL) and some rules for producing the correct word order in the output. It was soon recognised that specific dictionary=driven rules for syntactic ordering were too complex and increasingly ad hoc; the need for more systematic methods of syntactic analysis became evident. A number of projects were inspired by contemporary developments in linguistics, particularly Zellig Harris' and Noam Chomsky's ideas on syntactic transformations, but also other models such as dependency grammar and stratificational grammar. They seemed to offer the prospect of greatly improved translation.

Optimism remained at a high level for the first decade of MT research, with many predictions of imminent "breakthroughs", but disillusion grew as researchers encountered "semantic barriers" for which they saw no straightforward solutions. There were some operational systems - the Mark II system (developed by IBM and Washington University) installed at the USAF Foreign Technology Division, and the Georgetown University system at the US Atomic Energy Authority and at Euratom in Italy - but the quality of output was disappointing (although satisfying many recipients' needs for information). By 1964, the US government sponsors had become increasingly concerned at the lack of progress; they set up the Automatic Language Processing Advisory Committee (ALPAC), which concluded in its famous 1966 report that MT was slower, less accurate and twice as expensive as human translation and that "there is no immediate or predictable prospect of useful machine translation." It saw no need in the United States for further investment in MT research; instead it recommended the development of machine aids for translators, such as automatic dictionaries, and continued support in basic research in computational linguistics.

The ALPAC report was widely condemned as narrow, biased and shortsighted, but the damage had been done. It brought a virtual end of MT research in the United States for over a decade and it had great impact elsewhere in the Soviet Union and in Europe. However, MT research did continue in Canada, in France and in Germany. Within a few years Peter Toma, one of the members of the Georgetown University project, had developed Systran for operational use by the USAF (1970) and by NASA (in 1974/5), and shortly afterwards Systran was installed by the Commission of the European Communities for translating from English into French (1976) and later between other Community languages. At the same time, another successful operational system appeared in Canada, the METEO system for translating weather reports, developed at Montreal University.

In the 1960s in the US and the Soviet Union MT activity had concentrated on Russian-English and English-Russian translation of scientific and technical documents for a relatively small number of potential users, most of whom were prepared to overlook mistakes of terminology, grammar and style in order to be able to read something which they would have otherwise not known about. Since the mid-1970s the demand for MT has come from quite different sources with different needs and different languages. The administrative and commercial demands of multi-lingual communities and multinational trade have stimulated the demand for translation in Europe, Canada and Japan beyond the capacity of the traditional translation services. The demand is now for cost-effective machine-aided translation systems which can deal with commercial and technical documentation in the principal languages of international commerce.

The 1980s has witnessed the emergence of a variety of system types and from a widening number of countries. There are a number of mainframe systems. Best known is Systran, now installed worldwide and operating in many pairs of languages. Others are: Logos for German-English translation and for English-French in Canada; the internally developed systems for Spanish-English and English-Spanish translation at the Pan American Health Organization; the systems developed by the Smart Corporation for many large organizations in North America; and the recently marketed METAL system from Siemens for German-English translation. Major systems for English-Japanese and Japanese-English translation have come from Japanese computer companies, Fujitsu, Hitachi and Toshiba. The wide availability of microcomputers and of text-processing software has led to the commercial market for cheaper MT systems, exploited in North America and Europe by companies such as ALPS, Weidner, Linguistic Products, Tovna and Globalink, and by many Japanese companies, e.g. Sharp, NEC, Oki, Mitsubishi, Sanyo. Other microcomputer-based systems have appeared from China, Taiwan, Korea, Bolivia, Eastern Europe, the Soviet Union, etc.

Throughout the 1980s research on more advanced methods and techniques has continued. The dominant strategy is now that of 'indirect' translation via intermediary representations, sometimes interlingual in nature, involving semantic as well as morphological and syntactic analysis and sometimes non-linguistic 'knowledge bases'. There is increasing emphasis on devising systems for particular subject areas and particular specific purposes, for monolingual users as well as bilingual users (translators), and for interactive operation rather than batch processing. The most notable projects have been the GETA-Ariane system at Grenoble, SUSY and ASCOF at Saarbrücken, Mu at Kyoto, DLT at Utrecht, Rosetta at Eindhoven, the knowledge-based MT project at Carnegie-Mellon University (Pittsburgh), and two ambitious international multilingual projects: Eurotra, supported by the European Communities, involving teams in each member country; and the Japanese CICC project with participants in China, Indonesia and Thailand.

Linguistic problems of MT

The basic processes of translation are the analysis of the source language (SL) text, the conversion (or 'transfer') of the 'meaning' of the text into another language, and the generation (or 'synthesis') of the target language (TL) text. There are basically three overall strategies. In the 'direct translation' approach, adopted by most of the early MT projects, systems are designed in all details specifically for one particular pair of languages; vocabulary and syntax are not analysed any more than strictly necessary for the resolution of ambiguities, the identification of TL equivalents and output in correct TL word order; hence the processes of analysis and synthesis are combined in single programs, sometimes of monolithic intractability (e.g. the Georgetown system). The second strategy is the 'interlingua' approach which assumes the possibility of converting SL texts into (semantic) representations common to a number of languages, from which texts can be generated in one or more TLs. In interlingua systems SL analysis and TL synthesis are monolingual processes independent of any other languages, and the interlingua is designed to be language-independent or 'universal'. (A current example is the DLT system based on modified Esperanto representations.) The third strategy is the 'transfer' approach, which operates in three stages: from the SL text into an abstract 'intermediary' representation which is not language-independent but oriented to the characteristics of the SL (analysis); from such a SL-oriented representation to an equivalent TL-oriented representation (transfer); and from the latter to the final TL text (synthesis). (Major examples of the transfer approach are the GETA, SUSY, Mu and Eurotra systems.)

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 (*fly* as 'insect' or 'move through air', *bank* as 'edge of river' or 'financial institution') require different translations (*mouche*: *voler*; *rive*: *banque*); nouns can function as verbs (*control*, *plant*, *face*) and are hence 'ambiguous', since the TL may well have different forms (*contrôle*: *diriger*, *plante*: *planter*, *face*: *affronter*); other languages make distinctions which are absent in English: *river* can be French *rivière* or *fleuve*, German *Fluss* or *Strom*; *blue* can be Russian *sinii* or *goluboi*. Often, all combine, as illustrated by a simple but common example, the word *light*. This can be a noun meaning 'luminescence', an adjective meaning 'not dark', another adjective meaning 'not heavy', or a verb meaning 'to start burning' (at least). In French the meanings are conveyed by four different words *lumière*, *leger*, *clair*, *allumer*. An analysis of English must therefore distinguish the four possibilities by (a) recognising the grammatical categories of words in sentences (nouns, verbs, adjectives, adverbs, prepositions, conjunctions, etc.) and the structures in which they take part, and (b) by recognising the lexical and semantic contexts in which the words occur. At the transfer stage this information must be used to convert the identified meaning into those lexical units and structures with equivalent meanings in the target language. In many cases, differences between the vocabulary of the source and target languages are also accompanied by structural differences. A familiar example involves 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.

The choice of TL form involves a restructuring with effects on the translation of other lexical items (*what* as *ce que* and *wie*). A more radical, but no less common, instance of restructuring may be illustrated by the German sentence:

(3) Das Mädchen spielt gern Tennis

translated as:

(4) The girl likes to play tennis

The German adverb *gern* corresponds to an English finite verb *like*, and this choice entails the shifting of the finite verb *spielt* to a subordinate infinitive (*to play*).

The resolution of many linguistic problems transcends sentence boundaries. A common and persistently difficult one involves the use of pronouns. Following (3) might occur:

(5) Es geht jede Woche zum Club

for which the English should be:

(6) She goes to the club every week

However, *es* is normally translated as *it*. To ensure the correct selection of *she*, the preceding noun referent of the pronoun must be identified and the different practices for pronominalisation must be taken into account (in German according to the 'grammatical' gender of the preceding noun, and in English according to the 'natural' sex of the object referred to.) However, the identification of the noun referred to can often be more complex than this example. Frequently, it depends on (non-linguistic) knowledge of events or situations:

(7) The soldiers killed 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 'killing' implies 'death' and that 'death' is followed (normally) by 'burial'. This identification is crucial when translating into French where the pronoun must be *elles* and not *ils*. Grammatical and linguistic information is insufficient in such cases.

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: the predictive syntactic analyzer of the 1960s concentrated on sequences of categories (it was developed subsequently by Woods (1970) as the Augmented Transition Network parser); the dependency grammar (of Tesnière, Hays, etc.) has concentrated on functional relationships; and the phrase structure grammars have been the models for parsers of constituency structure. Each have their strengths and weaknesses, and modern MT systems often adopt an eclectic mixture of parsing techniques, now often within the framework of a 'unification grammar' formalism (Kay 1984).

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:

(8a) The camera was purchased by the man with dark glasses

(8b) The camera was purchased by the man with the tripod

(8c) The camera was purchased by the man with a cheque

A syntactic analysis can go no further than offer each possibility; later semantic or pragmatic analysis (e.g. involving lexical and situational context) has the task of specifying the intended relationship.

Many parsers now include the identification of case relations, e.g. the fact that in

(9) The house was built by a doctor for his son during the war.

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 *during the war*. 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 Attribute (8a), Comitative (8b), Instrument (8c). 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:

(10) He was beaten with a club

the 'social' sense of *club* found in (6) above is excluded by the verb-type which requires an 'inanimate' Instrument. In:

(11) The sailor went on board
(12) The sailor was examined by the board

the 'physical' sense of *board* in (11) is confirmed by the verb-type (motion) and the preposition of Location, and the 'social' sense in (12) is confirmed by the verb *examine* which requires an 'animate' Agent.

Few operational MT systems involve deeper levels of semantic or pragmatic analysis. Nevertheless, as examples (7) and (8) demonstrate, disambiguation and correct selection of TL equivalents would seem to be impossible without reference to knowledge of the features and properties of the objects and events described. This was used by Yehoshua Bar-Hillel (1960) in arguing that fully automatic translation of high quality is impossible. His famous demonstration involved the sentence *The box was in the pen*. We know that *pen* can refer here only to a 'container for animals or children' and not to 'writing implement', from our knowledge of relative sizes of (writing) pens and boxes. For Bar-Hillel, the incorporation of encyclopedic knowledge and the associated inference mechanisms was "utterly chimerical". However, subsequent advances in Artificial Intelligence (AI) have encouraged later MT researchers to investigate the possibility of knowledge-based systems (e.g. at Carnegie-Mellon University), at least for systems restricted to specific domains. However, the general feasibility of AI approaches has yet to be tested on large-scale systems, and most MT researchers prefer to develop linguistics-based systems capable of incorporating AI methods as adjuncts to more traditional techniques of syntactic and semantic analysis, transfer and generation.

MT in practice

The complexities and difficulties of linguistic analysis and the problems of incorporating appropriate semantic and extra-linguistic knowledge have persuaded many researchers that for the foreseeable future it is unrealistic to attempt to build fully automatic systems capable of the translation quality achieved by human translators. The growing demands for translations must be met by MT systems which involve the active assistance and expertise of natural language speakers.

The most obvious course, which has been adopted since the first MT systems, is to employ human translators to revise and improve the crude and inaccurate texts produced by MT systems. Initially 'post-editing' was undertaken manually; later systems incorporate on-line revision and in some cases special facilities for dealing with the most common types of error (e.g. transposition of words, insertion of articles). Revision for MT differs from the revision of traditionally produced translations; the computer program is regular and consistent with terminology, unlike the human translator, but typically it contains grammatical and stylistic errors which no human translator would commit.

The development of powerful microcomputer text editing facilities has led to the introduction of interactive MT systems. During the translation process, a human operator (normally a translator) may be asked to help the computer resolve ambiguities of vocabulary or structure, e.g. whether the *club* in (10) is a 'society' or not, and what relationship is expressed by *with* in (8a, 8b, and 8c). Many Japanese systems demand considerable assistance from operators, particularly with the 'pre-editing' of Japanese scripts (identifying word and phrase boundaries, punctuation, etc.)

A third possibility is to constrain the variety of language in the input texts. There are two approaches: either the system is designed to deal with one particular subject matter or the input texts are written in a vocabulary and style which it is known that the MT system can deal with. The former approach is illustrated by the METEO system, introduced in 1976, which translates weather forecasts from English into French for public broadcasts in Canada. The latter approach has been taken by the Xerox Corporation in its use of the Systran system; manuals are written in a controlled English (unambiguous vocabulary and restricted syntactic patterns) which can be translated with minimal revision into five languages. Other examples are the Smart systems installed at a number of large US and Canadian institutions which combine on-line editing to ensure clear documentation in English and 'restricted language' MT to produce translations for subsequent editing.

MT systems are now being used in the production of a wide range of translations of different quality and status. The 'raw' output of both mainframe systems (Systran, Logos, Fujitsu) and microcomputer systems (Weidner, NEC) may be used (a) as a draft version for full revision to the level of human quality products (e.g. for later publication), (b) as a first draft for subsequent wholly human translation, (c) as a version offered completely unedited to those who are prepared to tolerate the grammatical and stylistic errors for the sake of cheap access to information, or (d) as a version for light editing for similar information purposes. It may be noted, however, that few microcomputer-based translations are adequate in their unedited forms even for purely 'information' purposes.

A major and significant impact on the translation profession has been the development of computer-based aids for translators, which may justly be regarded as commercial by-products of research in MT and related areas. These aids include facilities for multilingual wordprocessing, for creating in-house glossaries and termbanks, for receiving and sending texts over telecommunication networks, for accessing remote sources of information, for publishing quality documents, and for using interactive or batch MT systems when appropriate. Systems which integrate various facilities of this nature are being developed as translator's workstations.

The languages of the earlier systems were mainly Russian and English, reflecting the political situation of the time. In the 1970s the main impetus was for systems to deal with the administrative needs of countries such as Canada and the European Communities, hence systems for English, French, German, and other Community languages. During the 1980s the main focus has been the languages of international trade and communications (English, Japanese, French, German, Spanish, and to lesser extent Chinese and Italian). On the other hand, the needs of Third World countries for scientific and technical textbooks in their own languages (mainly from English) are still not being fully met, although a start has been made by some individual projects (notably GETA) and by the Japanese multinational project.

The future of MT

In the immediate future, there will clearly be continued expansion and improvement of systems for the business and administrative communities. As at present, the MT market will include both microcomputer and mainframe systems. The cheaper microcomputer systems will produce relatively poor output needing substantial revision but which can be applied cost-effectively in commercial services. More expensive mainframe (or minicomputer) systems will be developed on transfer and interlingua approaches with some use of AI techniques. These will be producing higher quality output, which, although still requiring revision for many purposes (e.g. publication), will be satisfying basic information needs without revision.

It is probable that other types of systems will appear. Nearly all systems at present require users to know both source and target languages, generally to the level expected of regular translators. There is clearly a need for systems which can be used by individual non-translators ignorant of the source language in order to get translations giving at least the gist of document contents. At a further stage, these systems should be integrated with other documentation systems (information retrieval, abstracting, paraphrasing).

There is an equally clear need for systems for those ignorant of target languages, e.g. businessmen (and others) wanting to convey simple messages to make travel arrangements, to book hotel accommodation, to arrange meetings, etc. There has already been some recent research on 'interactive analysis' systems: the computer would seek to obtain from writers of texts information which would resolve ambiguities and which would enable the generation of appropriate texts. The interaction would be conducted in the user's own language.

Probably the most obvious area of future development will be speech translation. Research is already in progress (particularly in Japan) on systems for international telephone communication (initially restricted to standard business messages) which combine voice interpretation and voice production with machine translation. Given the problems of speech recognition in addition to the peculiarities of conversational language, operational prototypes are regarded very much as long-term objectives.

Nearly all developments depend on improvements in the automation of the basic translation processes. The ultimate ideal of fully automatic high quality translation may remain but it seems increasingly unrealistic. MT suffers still from what appear to be low-level problems: incorrect uses of pronouns, prepositions and tenses, erroneous translations of common vocabulary. Progress is slow, but developments in artificial intelligence, in linguistic theory, in computational linguistics and in computer technology promise future improvements in general quality.

At a more basic level much progress depends on the continued efforts to standardise terminology both within and across languages, which is of benefit to translators and technical writers generally. More specifically, the wasteful duplication involved in the creation of large MT dictionaries calls for inter-project cooperation, a process which has already started in Japan with the Electronic Dictionary Research project.

MT is already seen not as a threatening replacement of translators but as an aid to multilingual communication. The future development of MT rests on fruitful interaction between the researchers of experimental systems investigating new methods and theories, the developers of commercial systems exploiting well-tested methods in cost-effective practical systems, and the perception of the real needs of translators and other potential users of translation systems.

Further reading

For general introductions to MT see Lehrberger & Bourbeau (1988) or Hutchins & Somers (1992), for the general history of MT see Hutchins (1986); for descriptions of current systems and developments see Hutchins (1988), Slocum (1988) and Vasconcellos (1988).

References

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

Bar-Hillel, Y. (1960) 'The present status of automatic translation of languages' *Advances in Computers* 1: 91-163

Hutchins, W.J. (1986) *Machine translation: past, present, futu*re Chichester: Ellis Horwood. New York: Halsted Press.

Hutchins, W.J. (1988) 'Recent developments in machine translation: a review of the last five years.' In: *New directions in machine translation*, ed. D.Maxwell et al. (Dordrecht: Foris), 7-62

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

Kay, M. (1984) 'Functional unification grammar: a formalism for machine translation.' In: *Coling 84* (Stanford University), 75-78

Lehrberger, J. & Bourbeau, L. (1988) *Machine translation: linguistic characteristics of MT systems and general methodology of evaluation*. Amsterdam: Benjamins.

Slocum, J. (1988), ed. *Machine translation systems*. Cambridge: University Press.

Vasconcellos, M. (1988), ed. *Technology as translation strategy*. Binghampton, NY: State University of New York.

Wood, W. (1970) 'Transition network grammars for natural language analysis' *Communications of the ACM* 13: 591-606