1
General introduction and brief history

The mechanization of translation has been one of humanity's oldest dreams. In the twentieth century it has become a reality, in the form of computer programs capable of translating a wide variety of texts from one natural language into another. But, as ever, reality is not perfect. There are no 'translating machines' which, at the touch of a few buttons, can take any text in any language and produce a perfect translation in any other language without human intervention or assistance. That is an ideal for the distant future, if it is even achievable in principle, which many doubt.

What has been achieved is the development of programs which can produce 'raw' translations of texts in relatively well-defined subject domains, which can be revised to give good-quality translated texts at an economically viable rate or which in their unedited state can be read and understood by specialists in the subject for information purposes. In some cases, with appropriate controls on the language of the input texts, translations can be produced automatically that are of higher quality needing little or no revision.

These are solid achievements by what is now traditionally called Machine Translation (henceforth in this book, MT), but they have often been obscured and misunderstood. The public perception of MT is distorted by two extreme positions. On the one hand, there are those who are unconvinced that there is anything difficult about analysing language, since even young children are able to learn languages so easily; and who are convinced that anyone who knows a foreign

language must be able to translate with ease. Hence, they are unable to appreciate the difficulties of the task or how much has been achieved. On the other hand, there are those who believe that because automatic translation of Shakespeare, Goethe, Tolstoy and lesser literary authors is not feasible there is no role for any kind of computer-based translation. They are unable to evaluate the contribution which less than perfect translation could make either in their own work or in the general improvement of international communication.

1.1 The aims of MT

Most translation in the world is not of texts which have high literary and cultural status. The great majority of professional translators are employed to satisfy the huge and growing demand for translations of scientific and technical documents, commercial and business transactions, administrative memoranda, legal documentation, instruction manuals, agricultural and medical text books, industrial patents, publicity leaflets, newspaper reports, etc. Some of this work is challenging and difficult. But much of it is tedious and repetitive, while at the same time requiring accuracy and consistency. The demand for such translations is increasing at a rate far beyond the capacity of the translation profession. The assistance of a computer has clear and immediate attractions. The practical usefulness of an MT system is determined ultimately by the quality of its output. But what counts as a 'good' translation, whether produced by human or machine, is an extremely difficult concept to define precisely. Much depends on the particular circumstances in which it is made and the particular recipient for whom it is intended. Fidelity, accuracy, intelligibility, appropriate style and register are all criteria which can be applied, but they remain subjective judgements. What matters in practice, as far as MT is concerned, is how much has to be changed in order to bring output up to a standard acceptable to a human translator or reader. With such a slippery concept as translation, researchers and developers of MT systems can ultimately aspire only to producing translations which are 'useful' in particular situations — which obliges them to define clear research objectives — or, alternatively, they seek suitable applications of the 'translations' which in fact they are able to produce.

Nevertheless, there remains the higher ideal of equalling the best human translation. MT is part of a wider sphere of 'pure research' in computer-based natural language processing in Computational Linguistics and Artificial Intelligence, which explore the basic mechanisms of language and mind by modelling and simulation in computer programs. Research on MT is closely related to these efforts, adopting and applying both theoretical perspectives and operational techniques to translation processes, and in turn offering insights and solutions from its particular problems. In addition, MT can provide a 'test-bed' on a larger scale for theories and techniques developed by small-scale experiments in computational linguistics and artificial intelligence.

The major obstacles to translating by computer are, as they have always been, not computational but linguistic. They are the problems of lexical ambiguity, of syntactic complexity, of vocabulary differences between languages, of elliptical and 'ungrammatical' constructions, of, in brief, extracting the 'meaning' of sentences

and texts from analysis of written signs and producing sentences and texts in another set of linguistic symbols with an equivalent meaning. Consequently, MT should expect to rely heavily on advances in linguistic research, particularly those branches exhibiting high degrees of formalization, and indeed it has and will continue to do so. But MT cannot apply linguistic theories directly: linguists are concerned with explanations of the underlying 'mechanisms' of language production and comprehension, they concentrate on crucial features and do not attempt to describe or explain everything. MT systems, by contrast, must deal with actual texts. They must confront the full range of linguistic phenomena, the complexities of terminology, misspellings, neologisms, aspects of 'performance' which are not always the concern of abstract theoretical linguistics.

In brief, MT is not in itself an independent field of 'pure' research. It takes from linguistics, computer science, artificial intelligence, translation theory, any ideas, methods and techniques which may serve the development of improved systems. It is essentially 'applied' research, but a field which nevertheless has built up a substantial body of techniques and concepts which can, in turn, be applied in other areas of computer-based language processing.

1.2 Some preliminary definitions

The term Machine Translation (MT) is the now traditional and standard name for computerised systems responsible for the production of translations from one natural language into another, with or without human assistance. Earlier names such as 'mechanical translation' and 'automatic translation' are now rarely used in English; but their equivalents in other languages are still common (e.g. French *traduction automatique,* Russian *avtomatičeskii perevod).* The term does not include computer-based translation tools which support translators by providing access to dictionaries and remote terminology databases, facilitating the transmission and reception of machine-readable texts, or interacting with word processing, text editing or printing equipment. It does, however, include systems in which translators or other users assist computers in the production of translations, including various combinations of text preparation, on-line interactions and subsequent revisions of output. The boundaries between Machine-Aided Human Translation (MAHT) and Human-Aided Machine Translation (HAMT) are often uncertain and the term Computer-Aided (or Computer-Assisted) Translation (both CAT) can sometimes cover both. But the central core of MT itself is the automation of the full translation process.

Although the ideal may be to produce high-quality translations, in practice the output of most MT systems is revised (post-edited). In this respect, MT output is treated no differently than the output of most human translators which is normally revised by another translator before dissemination. However, the types of errors produced by MT systems do differ from those of human translators. While post-editing is the norm, there are certain circumstances when MT output may be left unedited (as a raw translation) or only lightly corrected, e.g. if it is intended only for specialists familiar with the subject of the text. Output may also serve as a rough draft for a human translator, as a pre-translation.

The translation quality of MT systems may be improved — not only, of course, by developing better methods — 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. polymer chemistry) and/or document type (e.g. patents). Alternatively, input texts may be written in a **controlled language,** which reduces potential ambiguities and restricts the complexity of sentence structures. This option is often referred to as **pre-editing,** but the term can also be used for the marking of input texts to indicate proper names, word divisions, prefixes, suffixes, phrase boundaries, etc. Finally the system itself may refer problems of ambiguity and selection to human operators (usually translators, though some systems are designed for use by the original authors) for resolution during the processes of translation itself, i.e. in an **interactive** mode.

Systems are designed either for one particular pair of languages **(bilingual** systems) or for more than two languages **(multilingual** systems), either in one direction only **(uni-directional** systems) or in both directions **(bi-directional** systems). In overall system design, there are three basic types. The first (and also historically oldest) is generally referred to as the **direct translation** approach: the MT system is designed in all details specifically for one particular pair of languages in one direction, e.g. Russian as the language of the original texts, the **source language,** and English as the language of the translated texts, the **target language.** Source texts are analysed no more than necessary for generating texts in the other language. The second basic type is the **interlingua** approach, which assumes the possibility of converting texts to and from 'meaning' representations common to more than one language. Translation is thus in two stages: from the source language to the interlingua, and from the interlingua into the target language. Programs for analysis are independent from programs for generation; in a multilingual configuration, any analysis program can be linked to any generation program. The third type is the less ambitious **transfer** approach. Rather than operating in two stages through a single interlingual meaning representation, there are three stages involving, usually, syntactic representations for both source and target texts. The first stage converts texts into intermediate representations in which ambiguities have been resolved irrespective of any other language. In the second stage these are converted into equivalent representations of the target language; and in the third stage, the final target texts are generated. Analysis and generation programs are specific for particular languages and independent of each other. Differences between languages, in vocabulary and structure, are handled in the intermediary transfer program.

Within the stages of analysis and generation, most MT system exhibit clearly separated components dealing with different levels of linguistic description: morphology, syntax, semantics. Hence, **analysis** may be divided into morphological analysis (e.g. identification of word endings), syntactic analysis (identification of phrase structures, etc.) and semantic analysis (resolution of lexical and structural ambiguities). Likewise, **generation** (or synthesis) may pass through levels of semantic, syntactic and morphological generation. In transfer systems, there may

be separate components dealing with lexical transfer (selection of vocabulary equivalents) and structural transfer (transformation of source text structures into equivalent target text ones).

In many older systems (particularly those of the direct translation type), rules for analysis, transfer and generation were not always clearly separated. Some also mixed linguistic data (dictionaries and grammars) and computer processing rules and routines. Later systems exhibit various degrees of modularity, so that system components, data and programs can be adapted and changed independently of each other.

1.3 Brief history of MT

The use of mechanical dictionaries to overcome barriers of language was first suggested in the 17th century. Both Descartes and Leibniz speculated on the creation of dictionaries based on universal numerical codes. Actual examples were published in the middle of the century by Cave Beck, Athanasius Kircher and Johann Becher. The inspiration was the 'universal language' movement, the idea of creating an unambiguous language based on logical principles and iconic symbols (as the Chinese characters were believed to be), with which all humanity could communicate without fear of misunderstanding. Most familiar is the interlingua elaborated by John Wilkins in his 'Essay towards a Real Character and a Philosophical Language' (1668).

In subsequent centuries there were many more proposals for international languages (with Esperanto as the best known), but few attempts to mechanize translation until the middle of this century. In 1933 two patents appeared independently in France and Russia. A French-Armenian, George Artsrouni, had designed a storage device on paper tape which could be used to find the equivalent of any word in another language; a prototype was apparently demonstrated in 1937. The proposal by the Russian, Petr Smirnov-Troyanskii, was in retrospect more significant. He envisaged three stages of mechanical translation: first, an editor knowing only the source language was to undertake the 'logical' analysis of words into their base forms and syntactic functions; secondly, a machine was to transform sequences of base forms and functions into equivalent sequences in the target language; finally, another editor knowing only the target language was to convert this output into the normal forms of that language. Although his patent referred only to the machine which would undertake the second stage, Troyanskii believed that "the process of logical analysis could itself be mechanised".

Troyanskii was ahead of his time and was unknown outside Russia when, within a few years of their invention, the possibility of using computers for translation was first discussed by Warren Weaver of the Rockefeller Foundation and Andrew D. Booth, a British crystallographer. On his return to Birkbeck College (London) Booth explored the mechanization of a bilingual dictionary and began collaboration with Richard H. Richens (Cambridge), who had independently been using punched cards to produce crude word-for-word translations of scientific abstracts. However, it was a memorandum from Weaver in July 1949 which

brought the idea of MT to general notice and suggested methods: the use of wartime cryptography techniques, statistical analysis, Shannon's information theory, and exploration of the underlying logic and universal features of language. Within a few years research had begun at a number of US centres, and in 1951 the first full-time researcher in MT was appointed: Yehoshua Bar-Hillel at MIT. A year later he convened the first MT conference, where the outlines of future research were already becoming clear. There were proposals for dealing with syntax, suggestions that texts should be written in controlled languages, arguments for the construction of sublanguage systems, and recognition of the need for human assistance (pre- and post-editing) until fully automatic translation could be achieved. For some, the first requirement was to demonstrate the technical feasibility of MT. Accordingly, at Georgetown University Leon Dostert collaborated with IBM on a project which resulted in the first public demonstration of a MT system in January 1954. A carefully selected sample of Russian sentences was translated into English, using a very restricted vocabulary of 250 words and just six grammar rules. Although it had little scientific value, it was sufficiently impressive to stimulate the large-scale funding of MT research in the United States and to inspire the initiation of MT projects elsewhere in the world, notably in the Soviet Union.

For the next decade many groups were active: some adopting empirical trial-and-error approaches, often statistics-based, with immediate working systems as the goal; others took theoretical approaches, involving fundamental linguistic research, aiming for long-term solutions. The contrasting methods were usually described at the time as 'brute-force' and 'perfectionist' respectively. Examples of the former were the lexicographic approach at the University of Washington (Seattle), later continued by IBM in a Russian-English system completed for the US Air Force, the statistical 'engineering' approach at the RAND Corporation, and the methods adopted at the Institute of Precision Mechanics in the Soviet Union, and the National Physical Laboratory in Great Britain. Largest of all was the group at Georgetown University, whose successful Russian-English system is now regarded as typical of this 'first generation' of MT research. Centres of theoretical research were at MIT, Harvard University, the University of Texas, the University of California at Berkeley, at the Institute of Linguistics in Moscow and the University of Leningrad, at the Cambridge Language Research Unit (CLRU), and at the universities of Milan and Grenoble. In contrast to the more pragmatically oriented groups where the 'direct translation' approach was the norm, some of the theoretical projects experimented with early versions of interlingua and transfer systems (e.g. CLRU and MIT, respectively).

Much of the research of this period was of lasting importance, not only for MT but also for computational linguistics and artificial intelligence — in particular, the development of automated dictionaries and of techniques for syntactic analysis — and many theoretical groups made significant contributions to linguistic theory. However, the basic objective of building systems capable of producing good translations was not achieved. Optimism had been high, there were many predictions of imminent breakthroughs, but disillusionment grew as the complexity of the linguistic problems became more and more apparent. In a 1960 review of MT progress, Bar-Hillel criticized the prevailing assumption that the goal of MT research should be the creation of fully automatic high-

quality translation (FAHQT) systems producing results indistinguishable from those of human translators. He argued that the 'semantic barriers' to MT could in principle only be overcome by the inclusion of vast amounts of encyclopaedic knowledge about the 'real world'. His recommendation was that MT should adopt less ambitious goals, it should build systems which made cost-effective use of human-machine interaction.

In 1964 the government sponsors of MT in the United States formed the Automatic Language Processing Advisory Committee (ALPAC) to examine the prospects. In its influential 1966 report it concluded that MT was slower, less accurate and twice as expensive as human translation and stated that "there is no immediate or predictable prospect of useful Machine Translation". It saw no need for further investment in MT research; instead it recommended the development of machine aids for translators, such as automatic dictionaries, and continued support of basic research in computational linguistics. The ALPAC report was widely condemned as narrow, biased and shortsighted — it was certainly wrong to criticize MT because output had to be post-edited, and it misjudged the economic factors — but large-scale financial support of current approaches could not continue. Its influence was profound, bringing a virtual end to MT research in the United States for over a decade and damaging the public perception of MT for many years afterwards.

In the following decade MT research took place largely outside the United States, in Canada and in Western Europe, and virtually ignored by the scientific community. American activity had concentrated on English translations of Russian scientific and technical materials. In Canada and Europe the needs were quite different: the Canadian bicultural policy created a demand for English-French (and to a less extent French-English) translation beyond the capacity of the market, and the European Economic Community (as it was then known) was demanding translations of scientific, technical, administrative and legal documentation from and into all the Community languages.

A research group was established at Montreal which, though ultimately unsuccessful in building a large English-French system for translating aircraft manuals, is now renowned for the creation in 1976 of the archetypal 'sublanguage' system Météo (Chapter 12) for translating weather reports for daily public broadcasting. In 1976 the Commission of the European Communities decided to install an English-French system called Systran, which had previously been developed by Peter Toma (once a member of the Georgetown team) for Russian-English translation for the US Air Force, and had been in operation since 1970 (see Chapter 10). In subsequent years, further systems for French-English, English-Italian, English-German and other pairs have been developed for the Commission. In the late 1970s, it was also decided to fund an ambitious research project to develop a multilingual system for all the Community languages, based on the latest advances in MT and in computational linguistics. This is the Eurotra project, which involves research groups in all member states (see Chapter 14).

For its basic design, Eurotra owes much to research at Grenoble and at Saarbrücken. During the 1960s the French group had built an 'interlingua' system for Russian-French translation (not purely interlingual as lexical transfer was still

bilingual); however, the results were disappointing and in the 1970s it began to develop the influential transfer-based Ariane system (Chapter 13). The Saarbrücken group had also been building its multilingual 'transfer' system SUSY since the late 1960s (Chapter 11). It was now the general consensus in the MT research community that the best prospects for significant advances lay in the development of transfer-based systems. The researchers at the Linguistics Research Center (LRC) at Austin, Texas (one of the few to continue after ALPAC) had come to similar conclusions after experimenting with an interlingua system and was now developing its transfer-based METAL system (Chapter 15); and in Japan work had begun at Kyoto University on the Mu transfer system for Japanese-English translation. The Eurotra group adopted the same basic approach, although it found subsequently that the demands of large-scale multilinguality led to the incorporation of many interlingual features.

However, during the 1980s the transfer-based design has been joined by new approaches to the interlingua idea. Most prominent is the research on knowledge-based systems, notably at Carnegie Mellon University, Pittsburgh (see section 18.1), which are founded on developments of natural language understanding systems within the Artificial Intelligence (AI) community. The argument is that MT must go beyond purely linguistic information (syntax and semantics); translation involves 'understanding' the content of texts and must refer to knowledge of the 'real world'. Such an approach implies translation via intermediate representations based on (extra-linguistic) 'universal' elements. Essentially non-AI-oriented interlingua approaches have also appeared in two Dutch projects: the DLT system at Utrecht based on a modification of Esperanto (Chapter 17) and the Rosetta system at Phillips (Eindhoven) which is experimenting with Montague semantics as the basis for an interlingua (Chapter 16)

More recently, yet other alternatives have emerged. For many years, automatic translation of speech was considered Utopian, but advances in speech recognition and speech production have encouraged the foundation of projects in Great Britain (British Telecom) and in Japan (Advanced Telecommunications Research, ATR): see section 18.6. The sophistication of the statistical techniques developed by speech research has revived interest in the application of such methods in MT systems; the principal group at present is at the IBM laboratories at Yorktown Heights, NY (see section 18.3)

The most significant development of the last decade, however, is the appearance of commercial MT systems. The American products from ALPSystems, Weidner and Logos were joined by many Japanese systems from computer companies (Fujitsu, Hitachi, Mitsubishi, NEC, Oki, Sanyo, Sharp, Toshiba), and in the later 1980s by Globalink, PC-Translator, Tovna and the METAL system developed by Siemens from earlier research at Austin, Texas. Many of these systems, particularly those for microcomputers, are fairly crude in the linguistic quality of their output but are capable of cost-effective operation in appropriate circumstances (see Chapter 9). As well as these commercial systems, there have been a number of in-house systems, e.g. the Spanish and English systems developed at the Pan-American Health Organization (Washington, DC), and the systems designed by the Smart Corporation for Citicorp, Ford, and the Canadian Department of Employment and Immigration. Many of the Systran installations

are tailor-made for particular organisations (Aérospatiale, Dornier, NATO, General Motors).

Nearly all these operational systems depend heavily on post-editing to produce acceptable translations. But pre-editing is also widespread: in some systems, for instance, operators are required, when inputting text, to mark word boundaries or even indicate the scope of phrases and clauses. At Xerox, texts for translation by Systran are composed in a controlled English vocabulary and syntax; and a major feature of the Smart systems is the pre-translation editor of English input.

The revival of MT research in the 1980s and the emergence of MT systems in the marketplace have led to growing public awareness of the importance of translation tools. There may still be many misconceptions about what has been achieved and what may be possible in the future, but the healthy state of MT is reflected in the multiplicity of system types and of research designs which are now being explored, many undreamt of when MT was first proposed in the 1940s. Further advances in computer technology, in Artificial Intelligence and in theoretical linguistics suggest possible future lines of investigation (see Chapter 18), while different MT user profiles (e.g. the writer who wants to compose a text in an unknown language) lead to new designs. But the most fundamental problems of computer-based translation are concerned not with technology but with language, meaning, understanding, and the social and cultural differences of human communication.

1.4 Further reading

The history of MT is covered by Hutchins (1986), updated by Hutchins (1988). Basic sources for the early period are Locke and Booth (1955) — which reproduces Weaver's memorandum — Booth (1967) and Bruderer (1982). For the period after ALPAC (1966) there are good descriptions of the major MT systems in King (1987), Nirenburg (1987a) and Slocum (1988), while Vasconcellos (1988) and the ongoing Aslib conference series *Translating and the Computer* (published under various editors) provide the wider perspectives of commercial developments and translators' experiences.