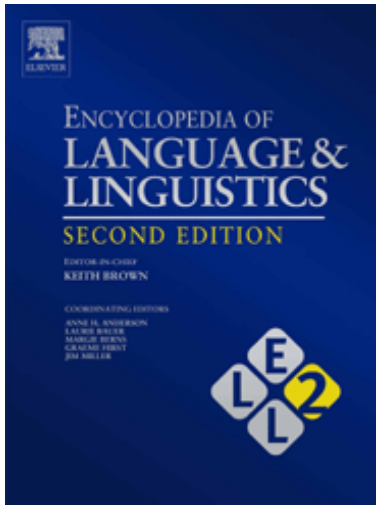




Hutchins J (2006), Machine Translation: History. In: Keith Brown, (Editor-in-Chief) *Encyclopedia of Language & Linguistics, Second Edition*, volume 7, pp. 375-383. Oxford: Elsevier.

Malmkjær K (2003). 'On a pseudo-subversive use of corpora in translator training.' In Zanettin F *et al.* (eds.). 119–134.
Mauranen A (2000). 'Strange strings in translated language: a study on corpora.' In Olohan M (ed.) *Intercultural faultlines*. Manchester: St. Jerome. 119–141.
Mauranen A (2002). 'Where's cultural adaptation?' *Intralinea 5*.
Mauranen A (2004). 'Corpora, universals and interference.' In Mauranen A & Kujamäki P (eds.). 65–82.
Mauranen A & Kujamäki P (eds.) (2004). *Translation universals: do they exist?* Amsterdam: Benjamins.
Olohan M (2001). 'Spelling out the optionals in translation: a corpus study.' *UCREL Technical Papers 13*, 423–432.
Olohan M & Baker M (2000). 'Reporting *that* in translated English: evidence for subconscious processes of explicitation?' *Across Languages and Cultures 1(2)*, 141–158.
Øverås L (1998). 'In search of the third code: an investigation of norms in literary translation.' *Meta 43(4)*, 571–588.
Pápai V (2004). 'Explicitation: a universal of translated text.' In Mauranen A & Kujamäkin P (eds.). 143–164.
Pearson J (2000). *Terms in context*. Amsterdam: Benjamins.
Pearson J (2003). 'Using parallel texts in the translator training environment.' In Zanettin *et al.* (eds.). 15–24.
Puurtinen T (1998). 'Syntax, readability and ideology in children's literature.' *Meta 43(4)*, 524–533.
Puurtinen T (2003). 'Genre-specific features of translationese? Linguistic differences between translated and non-translated Finnish children's literature.' *Literary and Linguistic Computing 18(4)*, 389–406.
Puurtinen T (2004). 'Explicitation of clausal relations: a corpus-based analysis of clause connectives in translated and non-translated Finnish children's literature.' In Mauranen A & Kujamäki P (eds.). 165–176.
Schmied J (2002). 'A translation corpus as a resource for translators: the case of English and German prepositions.' In Maia B, Haller J & Ulrych M (eds.) *Training the language services provider for the new millennium*. Porto: Universidade do Porto. 251–269.
Schmied J & Hildegard S (1996). 'Explicitness as a universal feature of translation.' In Ljung M (ed.) *Corpus-based studies in English: papers from ICAME 17*. Amsterdam/Atlanta: Rodopi. 21–36.
Scott M (1996). *Wordsmith Tools 3.0*. Oxford: Oxford University Press.
Teich E (2003). *Cross-linguistic variation in system and text*. The Hague: Mouton de Gruyter.
Tirkkonen-Condit S (2004). 'Unique items – over- or under-represented in translated language?' In Mauranen A & Kujamäki P (eds.). 177–184.
Tognini-Bonelli E (2001). *Corpus linguistics at work*. Amsterdam: Benjamins.
Tognini-Bonelli E & Manca E (2002). 'Welcoming children, pets and guests: a problem of nonequivalence in the languages of "agriturismo" and "farmhouse holidays".' *Textus 15(2)*, 317–334.
Toury G (1980). *In search of a theory of translation*. Tel Aviv: Tel Aviv University.
Toury G (1995). *Descriptive translation studies and beyond*. Amsterdam: Benjamins.
Toury G (2004). 'Probabilistic explanations in translation studies: welcome as they are, would they qualify as universals?' In Mauranen A & Kujamäki P (eds.). 15–32.
Vanderauwera R (1985). *Dutch novels translated into English: the transformation of a 'minority' literature*. Amsterdam: Rodopi.
Van Halteren H (ed.) (1999). *Syntactic Wordclass Tagging*. Dordrecht: Kluwer.
Varantola K (2003). 'Translators and disposable corpora.' In Zanettin F *et al.* (eds.). 55–70.
Zanettin F (1998). 'Bilingual comparable corpora and the training of translators.' *Meta 43(4)*, 616–630.
Zanettin F, Bernardini S & Stewart D (eds.) (2003). *Corpora in translator education*. Manchester: St. Jerome.

## Relevant Websites

http://www.wikipedia.org.



# Machine Translation: History

**J Hutchins**, Norwich, UK



## Precursors and Pioneers, 1933–1954

Although we might trace the origins of ideas related to machine translation (MT) to 17th-century speculations about universal languages and mechanical dictionaries, it was not until the 20th century that the first practical suggestions could be made, in 1933 with two patents issued in France and Russia to Georges Artsrouni and Petr Trojanskij, respectively. Artsrouni's patent was for a general-purpose machine that could also function as a mechanical multilingual dictionary. Trojanskij's patent, also basically for a mechanical dictionary, went further with detailed proposals for coding and interpreting grammatical functions using 'universal' (Esperanto-based) symbols in a multilingual translation device.

Neither of these precursors was known to Andrew Booth (a British crystallographer) and Warren Weaver

(a director at the Rockefeller Foundation) when they met in 1946 and 1947 and discussed tentative ideas for using the newly invented computers for translating natural languages. In July 1949, a memorandum by Weaver, which stimulated the start of MT research in the United States, suggested various methods based on his knowledge of cryptography, statistics, information theory, logic, and language universals (Hutchins, 1997).

In June 1952, the first MT conference was convened at MIT by Yehoshua Bar-Hillel, who had been appointed to survey the field. It was already clear that full automation of good-quality translation was a virtual impossibility and that human intervention either before or after computer processes (known from the beginning as pre- and post-editing, respectively) would be essential. Ideas were put forward for normalizing input texts and for microglossaries to reduce ambiguity problems and for some kind of syntactic structure analysis. Suggestions for future activity were proposed; in particular, Léon Dostert from Georgetown University argued for a public demonstration of the feasibility of MT.

Accordingly, he collaborated with IBM on a simple MT system demonstrated in New York on January 7, 1954, with a great deal of media attention. 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 the system had little scientific value, its output was sufficiently impressive to stimulate large-scale funding in the United States and to inspire the initiation of MT projects elsewhere, notably in the USSR.

## High Expectations, 1954–1966

When MT research began, there was little help to be had from current linguistics. As a consequence, in the 1950s and 1960s, researchers tended to polarize between empirical trial-and-error approaches, often using statistical methods to discover grammatical and lexical regularities that could be applied computationally, and theoretical approaches involving fundamental linguistic research and indeed the beginnings of what was later called computational linguistics (*see* **Computational Linguistics: History**). This decade saw the beginnings of three basic approaches to MT (*see* **Machine Translation: Overview**): (1) the direct translation model, in which programming rules were developed for the translation specifically from one source language (SL) into one particular target language (TL) with a minimal amount of analysis and syntactic reorganization and in which problems of homonymy and ambiguity were simplified by providing single TL equivalents for SL words to cover most senses; (2) the interlingua model (*see* **Machine Translation: Interlingual Methods**), in which translation was into and from some kind of language-neutral representation and that involved complex syntactic and semantic analysis; and (3) the transfer model, in which SL texts were analyzed into disambiguated SL representations and then converted into equivalent TL representations, from which output text was generated.

The direct translation approach was epitomized by research by Erwin Reifler (at the University of Washington, Seattle) and by Gilbert King (at IBM). Large bilingual dictionaries were compiled for a photoscopic store (a purpose-built memory device), in which lexicographic information was used not only for selecting lexical equivalents but also for solving grammatical problems without the use of syntactic analysis. A Russian-English system was installed for the U.S. Air Force, producing translations until the early 1970s – the output was crude and barely intelligible, but appeared to satisfy basic information needs of users.

Dictionary development was also the focus of research at a number of other centers. Anthony Oettinger's group (Harvard University) compiled a massive Russian-English dictionary, both to serve as an aid for translators (a forerunner of the now-common computer-based dictionary aids) and to produce crude word-for-word translations. Sydney Lamb (at the University of California, Berkeley) concentrated on developing maximally efficient dictionary routines and a linguistic theory appropriate for MT – his theory of stratificational grammar.

The system developed at Georgetown University, for many years the largest research group in the United States, was also essentially direct translation in approach, although incorporating various levels of structure analysis: morphological, syntagmatic (noun-adjective agreement, verb government, etc.), and syntactic (subjects and predicates, clause relationships, etc.). Systems installed by Euratom in 1963 and by the U.S. Atomic Energy Commission in 1964 continued in regular use until the late 1970s. A direct translation approach similar to the Georgetown system was made at the Institute of Precision Mechanics (under D.Y. Panov), but with less practical success, primarily because of lack of adequate computer facilities.

Research on interlinguas was almost wholly theoretical. The Cambridge Language Research Unit (Margaret Masterman) investigated a prototype interlingua to produce crude (almost word-for-word) translations, which would be refined by means of the semantic networks of a thesaurus (using mathematical

lattice theory.) Silvio Ceccato (Milan University) developed an interlingua based on cognitive processes, involving the conceptual analysis of words and their possible correlations in texts – a forerunner of the neural networks of later years. Igor A. Mel'čuk (Institute of Linguistics, Moscow) worked on the linguistic foundations of an interlingua, his stratificational dependency (meaning-text) model of language (*see* **Mel'čuk, Igor Aleksandrovic (b.1932)**). Nikolaj Andreev (Leningrad State University) conceived an interlingua as composed of those features statistically most common in a large number of languages.

The transfer approach was epitomized by the research at MIT led by Victor Yngve on syntactic analysis and production, mainly for German and English. Despite Chomsky's association with the group for a short time, transformational grammar (*see* **Generative Grammar**) had little influence – indeed Chomskyan linguistics had little impact on any MT group in this period.

Apart from MIT, the most explicit concentration on syntactic issues was at Harvard (after 1959), where research focused on the predictive syntactic analyzer (originally developed at the National Bureau of Standards under Ida Rhodes), a system for the identification of permissible sequences of grammatical categories (nouns, verbs, adjectives, etc.) and the probabilistic prediction of following categories. However, the results were often unsatisfactory, caused primarily by the enforced selection at every stage of the 'most probable' prediction. (Nevertheless, an improved version, the Multiple-Path Predictive Analyzer, led later to William Woods's familiar Augmented Transition Network parser.)

Computer facilities were frequently inadequate, and much effort was devoted to improving basic hardware (paper tapes, magnetic media, access speeds, etc.) and to devising programming tools suitable for language processing (e.g., COMIT developed at MIT). Some groups were inevitably forced to concentrate on theoretical issues, particularly in Europe and the Soviet Union. For political and military reasons, nearly all U.S. research was for Russian-English translation, and most Soviet research focused on English-Russian systems, although the multilingual policy of the Soviet Union inspired research there on a much wider range of languages than elsewhere.

By the mid-1960s MT research groups had been established in many countries throughout the world, including most European countries (Hungary, Czechoslovakia, Bulgaria, Belgium, Germany, France, etc.), China, Mexico, and Japan. Many of these were short-lived; an exception was the project that begun in 1960 at Grenoble University.

Throughout this period, research on MT became an umbrella for much contemporary work in structural and formal linguistics (particularly in the Soviet Union), semiotics, logical semantics, mathematical linguistics, quantitative linguistics, and nearly all of what is now called computational linguistics and language engineering – and in the Soviet Union with close ties with cybernetics and information theory (Léon, 1997).

## The ALPAC Report and Its Consequences

In the 1950s, optimism was high; developments in computing and in formal linguistics, particularly in the area of syntax, seemed to promise great improvements in quality; there were many predictions of fully automatic systems operating within a few years. However, disillusion grew as the complexity of the linguistic problems became more apparent and research seemed to face an apparently insuperable semantic barrier. In an influential survey, Bar-Hillel (1960) 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 it was not merely unrealistic, given the current state of linguistic knowledge and computer systems, but impossible in principle.

Although the first working systems (from IBM and Georgetown) were showing that poor-quality translations could be useful in many circumstances, there was growing disappointment at the slow development of good-quality MT. In 1964, the government sponsors of MT in the United States (mainly military and intelligence agencies) asked the National Science Foundation to set up the Automatic Language Processing Advisory Committee (ALPAC) to examine the situation. It concluded 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" (ALPAC, 1966: 32). 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 the continued support of basic research in computational linguistics. Paradoxically, ALPAC rejected MT because it fell well short of human quality (i.e., it required post-editing) even though human translations are invariably revised before publication and even though the sponsoring bodies were primarily interested in information gathering and analysis, in which lower quality would be acceptable. The influence of ALPAC was profound, bringing virtually to an end MT research in the United States for over a

decade and indirectly bringing to an end much MT research elsewhere. As a consequence, MT was to be no longer the leading area of research in computers and natural language.

## The Quiet Decade, 1967–1976

Research did not stop completely, however. Even in the United States, groups continued for a few more years at the University of Texas and at Wayne State University. But there was a change of direction. In the United States, activity had concentrated on English translations of Russian scientific and technical materials. In Canada and Europe, the needs were quite different. The Canadian government's bicultural policy created a demand for English-French (and to a lesser extent French-English) translation beyond the capacity of the translation profession. The demand was even greater within the offices of the European Community for translations of documents from and into all the EC languages.

At Montreal, research began in 1970 on a syntactic transfer system for English-French translation. The TAUM project (Traduction Automatique de l'Université de Montréal) had two major achievements: first, the Q-system formalism for manipulating linguistic strings and trees (later developed as the Prolog programming language) and, second, the Météo system designed specifically for translating the restricted vocabulary and limited syntax (sublanguage) of meteorological reports, which were publicly broadcast from 1976.

Other groups continued with essentially interlingua approaches. In the Soviet Union, Mel'čuk continued his research on a meaning-text model for application in MT. At Grenoble University, Bernard Vauquois developed a pivot language for translating Russian mathematics and physics texts into French – not, however, a full interlingua because it represented only the logical properties of syntactic relationships; there were no interlingual representations for lexical items, which were translated by a bilingual transfer mechanism. A similar model was adopted at the University of Texas (under Winfred Lehmann) in its METAL system for German and English: Sentences were analyzed into 'normal forms', that is, interlingual semantic propositional dependency structures with no interlingual lexical elements.

By the mid-1970s, however, the Grenoble and Texas groups recognized major problems with their interlingua approaches: the rigidity of the levels of analysis (failure at any one stage meant failure to produce any output at all), the inefficiency of parsers (too many partial analyses that had to be filtered out), and in particular loss of information about surface forms of SL input that might be used to guide selection of TL forms and construction of acceptable TL sentence structures. To many at the time it seemed that the less ambitious transfer approach offered better prospects.

## The Revival, 1976–1989

In the decade after ALPAC, more MT systems came into operational use and attracted public attention. Most significant were the first Systran installations. Developed by Peter Toma, its oldest version is the Russian-English system at the USAF Foreign Technology Division (Dayton, Ohio) installed in 1970. The Commission of the European Communities installed its first Systran system in 1976 and now has versions for most languages of the European Community (now European Union). Over the years, the original (basically, direct translation) design has been greatly modified, with increased modularity and greater compatibility of the analysis and synthesis components of different versions, permitting cost reductions when developing new language pairs. During the 1980s and subsequently, Systran was also installed at many major companies (e.g., General Motors of Canada, Dornier, and Aérospatiale).

From the early 1980s until the mid-1990s, Systran's main commercial rival was the Logos Corporation. After experience with an English-Vietnamese system for translating aircraft manuals during the 1970s, Logos developed a successful German-English system that first appeared on the market in 1982; during the 1980s, other language pairs were also developed. Also in the 1980s a commercial METAL German-English system appeared, the research at the University of Texas University now funded by Siemens of Munich (Germany). The system was no longer interlingua-based but was now essentially a transfer-based approach. Other language pairs were later marketed for Dutch, French, and Spanish, as well as for English and German.

Research also revived in the late 1970s and early 1980s. In contrast to the focus in the late 1960s and early 1970s on interlingua approaches, this was characterized by the widespread adoption of the three-stage transfer-based approach, predominantly syntax-oriented and founded on the formalization of lexical and grammatical rules influenced by the linguistic theories of the time.

One major exemplar was the system developed at Grenoble (GETA, Groupe d'Etudes pour la Traduction Automatique), the influential Ariane system. Regarded as the paradigm of the second-generation linguistics-based transfer systems, Ariane influenced projects throughout the world in the 1980s.

Of particular note were its flexibility and modularity and its algorithms for manipulating tree representations that incorporated many different levels and types of representation (dependency, phrase structure, and logical). Similar in conception were the multilingual transfer system at Saarbrücken (called SUSY), incorporating a heterogeneity of techniques (phrase structure rules, transformational rules, case grammar and valency frames, dependency grammar, the use of statistical data, etc.), and the Mu system developed at the University of Kyoto under Makoto Nagao, where the most prominent features were the use of case grammar analysis and dependency tree representations. (In 1986, the research prototype was converted to an operational system for use by the Japanese Information Center for Science and Technology for the translation of abstracts.)

One of the best-known projects of the 1980s was the Eurotra project of the European Community, intended as an advanced multilingual transfer system for translation among all the EC languages. Like GETA-Ariane and SUSY, the design combined lexical, logico-syntactic, and semantic information in multilevel interfaces at a high degree of abstractness. No facilities for human assistance during translation processes were envisaged, and problems of the lexicon were also neglected; at the end of the 1980s with no operational system in prospect, the project ended, having, however, achieved its secondary aim of stimulating cross-national research in computational linguistics.

From the mid-1980s, there was a revival of interest in interlingua systems, motivated in part by contemporary research in artificial intelligence and cognitive linguistics. Two major projects were based in the Netherlands. The DLT (Distributed Language Translation) system in Utrecht (under Toon Witkam) was intended as a multilingual interactive system operating over computer networks, in which each terminal was to be a translating machine from and into one language only, using a modified form of Esperanto as an interlingua. The project made a significant effort in the construction of large lexical databases and proposed the use of a Bilingual Knowledge Bank from a corpus of (human) translated texts (anticipating later example-based approaches). The Rosetta project at Eindhoven (under Jan Landsbergen) explored the use of Montague grammar (*see* **Montague Semantics**) – semantic interlingual representations were derived from the interpretation of the derivation trees of syntactic representations – and the feasibility of reversible grammars for analysis and generation. Reversibility subsequently became a feature of many MT projects.

In the latter half of the 1980s, Japan witnessed a substantial increase in activity. Most of the computer companies (Fujitsu, Toshiba, Hitachi, etc.) began to invest large sums into areas that government and industry saw as fundamental to the coming fifth generation of the information society – this included MT. The research, initially greatly influenced by the Mu project at Kyoto University, showed a wide variety of approaches. Although transfer systems predominated, there were also interlingua systems, such as the PIVOT system at NEC and the Japanese-funded multilingual multinational project, with participants from China, Indonesia, Malaysia, and Thailand. There was also considerable commercial activity and most of the computer companies marketed translation software, mainly for the Japanese-English and English-Japanese markets. Many of these systems were low-level direct translation or transfer systems limited to morphological and syntactic analysis and often with little attempt made to automatically resolve lexical ambiguities. Often restricted to specific subject fields (computer science and information technology were popular choices), they relied on substantial human assistance at both the preparatory (pre-editing) and the revision (post-editing) stages.

During the 1980s, many research projects were also established in Korea, in Taiwan, in mainland China, and in Southeast Asia, particularly in Malaysia. There was also an increase in activity in the Soviet Union, where from 1976 research was concentrated at the All-Union Center for Translation in Moscow. Systems for English-Russian and German-Russian translation were developed based on the direct translation approach, but there was also work under the direction of Yurij Apres'jan, based on Mel'čuk's meaning-text model, leading to systems for French-Russian and English-Russian. Apart from this group, however, most activity in the Soviet Union focused on the production of relatively low-level operational systems, often involving the use of statistical analyses (influenced by the Speech Statistics group under Rajmund Piotrovskij at Leningrad State University).

During this period, many researchers believed that quality improvements would come from language research in the context of artificial intelligence (AI; *see* **Natural Language Understanding, Automatic**), such as the investigations by Yorick Wilks on preference semantics and semantic templates and by Roger Schank on expert systems and knowledge-based text understanding. The Japanese investment in artificial intelligence projects also had a major impact and may well have stimulated the revival of government funding in the United States. A number of projects applied

knowledge-based approaches in Japan, in Europe, and particularly in North America. In the United States, the most important was at Carnegie-Mellon University (Pittsburgh) under Jaime Carbonell and Sergei Nirenburg, which experimented with a number of knowledge-based MT systems (Nirenburg *et al*., 1992).

For syntactic processing, there was a trend toward the adoption of unification and constraint-based formalisms (e.g., Lexical-Functional Grammar, Head-Driven Phrase Structure Grammar; Categorial Grammar, etc.). Complex multilevel representations and large sets of transformation and mapping rules were replaced by lexicalist approaches: monostratal representations, a restricted set of abstract rules, and constraints incorporated into specific lexical entries.

As far as practical use was concerned, one method of improving quality was seen to be the 'control' of vocabulary and syntax. Systems such as Systran, Logos, and METAL were in principle designed for general application, although in practice their dictionaries are adapted for particular subject domains by corporation purchasers. The Xerox Corporation went further by restricting, or controlling, input texts, so that transfer and generation was simpler and there was less need for post-editing. They have been followed by many others in subsequent years – indeed, the design of controlled languages is itself an active area of research (*see* **Controlled Languages**). There have been several companies that design controlled language systems for specific clients (e.g., LANT (later Xplanation) and ESTeam). The major example since the early 1980s has been the Smart Corporation (New York), with customers such as Citicorp, Ford, and the Canadian Department of Employment and Immigration. Restriction to particular subject areas does not necessarily mean controlled input. Special-purpose systems can be designed for particular environments, allowing for less complication in lexicons. The main examples here are the successful Spanish-English (SPANAM) and English-Spanish (ENGSPAN) systems developed during the 1970s and 1980s at the Pan American Health Organization in Washington (under Muriel Vasconcellos).

## Developments since 1989

The dominant framework for MT research until the end of the 1980s was essentially based on linguistic rules for syntactic analysis, for lexical transfer, for morphology, and so forth. Since 1989, however, this dominance has been broken by the emergence of new methods and strategies loosely called corpus-based methods.

Most dramatic has been the revival of statistics-based approaches – seen as a return to the empiricism of the first decade as opposed to the rationalism of the later rule-based methods. The two sides, empiricists and rationalists, were first directly confronted at a conference in Montreal in June 1992 (TMI, 1992).

With the success of their stochastic techniques in speech recognition, a group at IBM (Yorktown Heights, New York) developed a statistical MT (SMT) model, Candide (Brown *et al*., 1990). Its distinctive feature was that no linguistic rules were applied; statistical methods were virtually the sole means of analysis and generation. Candide was applied to the corpus of French and English texts in the reports of Canadian parliamentary debates. The first step was the alignment of SL and TL sentences and words that are potentially translation equivalents. Translation itself was achieved through a translation model (frequency statistics derived from previously aligned SL and TL words and phrases) and a language model of TL word sequences. During the 1990s, statistical methods (*see* **Language Processing: Statistical Methods**) have been the main focus of many research groups, for example, the improvement of bilingual text and word alignment techniques and statistical methods for extracting lexical and terminological data from bilingual corpora. Some have concentrated on the development of purely statistics-based systems, statistical MT (e.g., at the universities of Southern California, Aachen, and Hong Kong), and others have investigated the integration of statistical and traditional rule-based approaches (e.g., at Carnegie-Mellon University; the National Tsing-Hua University, Taiwan; and Microsoft Corporation).

The second major corpus-based approach, benefiting likewise from improved rapid access to large databanks of text corpora, has been what is known as the example-based (or memory-based) approach. Although first proposed in 1981 by Makoto Nagao, it was only toward the end of the 1980s that experiments began, initially in some Japanese groups and during the DLT project. The underlying hypothesis is that translation often involves the finding (or recalling) of analogous examples, that is, how a particular expression or similar phrase has been translated before. Example-based MT (EBMT) involves the matching of input phrases against equivalent or similar SL phrases in a corpus of parallel bilingual texts (aligned either statistically or by traditional rule-based methods), the extraction of TL equivalent or closely matching phrases (using, e.g., semantic networks or lexical frequencies), and – the most difficult process – the 're-combination' of selected TL phrases to produce fluent and grammatical output.

Example-based MT research is pursued actively at a number of Japanese and U.S. centers (Carl and Way, 2003).

Although the main innovations since 1990 have been corpus-based approaches, rule-based research has continued with projects on both transfer and interlingua systems. Examples of transfer systems are the PaTrans system for Danish/English translation of patents (based on Eurotra research) and the LMT system of Michael McCord at IBM. Interlingua-based projects are the CATALYST knowledge-based system at Carnegie-Mellon University for the Caterpillar company and the special-purpose systems (DIPLOMAT) developed for military operations, also at Carnegie-Mellon; the ULTRA system at New Mexico State University (Sergei Nirenburg); the UNITRAN system based on the linguistic theory of Principles and Parameters (*see* **Principles and Parameters Framework of Generative Grammar**) (Bonnie J. Dorr, University of Maryland); the Pangloss project, a collaborative project involving the universities of Southern California, New Mexico State, and Carnegie-Mellon; and the Universal Networking Language project at the Institute of Advanced Studies of the United Nations University (Tokyo), involving groups in some 15 countries.

Another new departure for MT research in the 1990s has been the growing interest in spoken language translation, with the challenge of combining speech recognition and the interpretation and production of conversation and dialog. The first long-standing group was established in 1986 at ATR Interpreting Telecommunications Research Laboratories (Nara, Japan), which has been developing a system for telephone registrations at international conferences and for telephone booking of hotel accommodation (Kurematsu and Morimoto, 1996). Slightly later came the JANUS project (under Alex Waibel) in a consortium of Carnegie-Mellon University, the University of Karlsruhe, and ATR. The JANUS project has also focused on travel planning, but the system is designed to be readily expandable to other domains (Levin *et al*., 2000). Both projects continue. A third shorter-lived group, by SRI (Cambridge, UK) as part of its Core Language project, investigated Swedish-English translation via quasi-logical forms (Rayner *et al*., 2000); and on a larger scale there was a fourth project, Verbmobil (directed by Wolfgang Wahlster), funded from 1993 until 2000 by the German government. Its aim was to develop a transportable aid for face-to-face English-language commercial negotiations by Germans and Japanese not speaking fluent English. Although the basic goal was not achieved, efficient methodologies for dialog and speech translation were developed and the project established top-class research groups in German universities as a valuable by-product (Wahlster, 2000).

Evaluation of MT systems emerged as a major focus of research activity, now recognized as crucial for progress in MT research itself. Since the early 1990s, there have been numerous workshops dedicated specifically to problems of evaluation (often attached to MT conferences), and in particular there has been the series of Language Resources and Evaluation Conferences (LREC). Methodologies developed by the Japan Electronic Industry Development Association (JEIDA) and the Expert Advisory Group on Language Engineering Standards (EAGLES) and for the evaluation of ARPA supported projects have been particularly influential.

The use of MT systems expanded greatly in the 1990s, particularly in commercial agencies, government services, and multinational companies, where translations are produced on a large scale, primarily of technical documentation. This was and remains the major market for the mainframe systems (Systran, Logos, METAL, and ATLAS), now usually on client-server configurations. Already in 1995, it was estimated that over 300 million words a year were being translated by such companies.

The most significant practical development for human translators has been the appearance in the early 1990s of the first translation workstations, which combine various machine aids (*see* **Machine-aided Translation, Methods**): multilingual word processing, OCR facilities, terminology management software, facilities for concordancing (facilities that translators had become familiar with in the 1980s ), – and in particular translation memories. The historical origins are described in Hutchins (1998).

Although MT systems for personal computers began to be marketed in the 1980s (e.g., the systems from Weidner, Globalink, and Toshiba), there has been a great expansion since 1990. The increasing computational power and storage capacities of personal computers makes these commercial systems the equal of previous mainframe systems of the 1980s and earlier – and, in many cases, more powerful. However, there has not been a matching improvement in translation quality. Nearly all are based on older transfer-based (or even direct translation) models; few have substantial and well-founded dictionaries; and most attempt to function as general-purpose systems, although most vendors do offer specialist dictionaries. In nearly all cases, systems are sold in three basic versions: systems for large corporations (enterprise systems), usually running

on client-server configurations; systems intended for independent translators (professional systems); and systems for nontranslators (home use).

The Internet has had a major impact since the mid-1990s. First, there has been the appearance of MT software products specifically for translating Web pages and electronic mail messages. Second, beginning in the mid-1990s, many MT vendors have provided Internet-based online translation services for translation on demand – pioneered by Systran on the French Minitel network during the 1980s, by CompuServe in 1995 based on the Transcend system, and then shortly afterward by AltaVista (the Babelfish service using Systran). There are now numerous other online services, some offering post-editing by human translators (revisers), at extra cost, but in most cases presenting unrevised results. Hence, translation quality is often poor, inevitably given the colloquial nature of many source texts, but these services are undoubtedly filling a significant (and apparently widely acceptable) demand for immediate rough translations for information purposes. MT is now reaching a mass market.

## Further Reading

The general history of MT is covered by Hutchins (1986), updated by Hutchins (1988, 1993). Basic sources for the early period are Locke and Booth (1955), Edmundson (1961), Booth (1967), Rozencvejg (1974), Henisz-Dostert *et al*. (1979), Bruderer (1982), and Hutchins (2000). For the 1970s and 1980s, there are good descriptions of the main systems in Nirenburg (1987), King (1987), and Slocum (1988). For systems developed during the 1990s, sources include Dorr *et al*. (1999); Somers (2003), the journal *Machine Translation*, the biennial Machine Translation Summit conferences, workshops and other conferences for MT, conferences for language resources and evaluation, computational linguistics, and the Machine Translation Archive.

*See also:* Chomsky, Noam (b. 1928); Computational Linguistics: History; Controlled Languages; Generative Grammar; Language Processing: Statistical Methods; Machine Translation: Interlingual Methods; Machine Translation: Overview; Machine-Aided Translation: Methods; Mel'čuk, Igor Aleksandrovic (b.1932); Montague Semantics; Natural Language Understanding, Automatic; Parsing and Grammar Description, Corpus-Based; Principles and Parameters Framework of Generative Grammar; Speech Acts and Artificial Intelligence Planning Theory.

## Bibliography

ALPAC (1966). *Language and machines: computers in translation and linguistics*. Report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council. Washington, DC: National Academy of Sciences, National Research Council.

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

Booth A D (ed.) (1967). *Machine translation*. Amsterdam: North-Holland.

Brown P F, Cocke J, Della Pietra S A *et al*. (1990). 'A statistical approach to machine translation.' *Computational Linguistics 16(2)*, 79–85.

Bruderer H E (ed.) (1982). *Automatische Sprachübersetzung*. Darmstadt: Wissenschaftliche Buch-Gesellschaft.

Carl M & Way A (eds.) (2003). *Recent advances in example-based machine translation*. Dordrecht: Kluwer Academic Publishers.

Dorr B J, Jordan P W & Benoit J W (1999). 'A survey of current paradigms in machine translation.' *Advances in Computers 49*, 1–68.

Edmundson H P (ed.) (1961). *Proceedings of the National Symposium on Machine Translation*. London: Prentice-Hall.

Henisz-Dostert B, Macdonald R R & Zarechnak M (1979). *Machine translation*. The Hague: Mouton.

Hutchins W J (1986). *Machine translation: past, present, future*. Chichester, UK: Ellis Horwood/New York: John Wiley.

Hutchins W J (1988). 'Recent developments in machine translation: a review of the last five years.' In Maxwell D *et al*. (eds.) *New directions in machine translation*. Dordrecht: Foris. 9–63.

Hutchins W J (1993). 'Latest developments in machine translation technology: beginning a new era in MT research.' In *MT Summit IV: international cooperation for global communication*. Tokyo: AAMT. 11–34.

Hutchins W J (1997). 'From first conception to first demonstration: the nascent years of machine translation, 1947–1954. A chronology.' *Machine Translation 12(3)*, 195–252.

Hutchins W J (1998). 'The origins of the translator's workstation.' *Machine Translation 13(4)*, 287–307.

Hutchins W J (ed.) (2000). *Early years in machine translation: memoirs and biographies of pioneers*. Amsterdam/Philadelphia: John Benjamins.

King M (ed.) (1987). *Machine translation today: the state of the art*. Edinburgh, UK: Edinburgh University Press.

Kurematsu A & Morimoto T (1996). *Automatic speech translation: fundamental technology for future cross-language communications*. Amsterdam: Gordon and Breach.

Léon J (1997). 'Les premières machines à traduire (1948–1960) et la filiation cybernétique.' *Bulag 22*, 9–33.

Levin L, Lavie A, Woszczina M, Gates D *et al.* (2000). 'The JANUS-III translation system: speech-to-speech translation in multiple domains.' *Machine Translation 15(1–2),* 3–25.
Locke W N & Booth A D (eds.) (1955). *Machine translation of languages: fourteen essays*. Cambridge, MA: MIT Technology Press.
Nirenburg S (ed.) (1987). *Machine translation: theoretical and methodological issues*. Cambridge, UK: Cambridge University Press.
Nirenburg S, Carbonell J, Tomita M & Goodman K (1992). *Machine translation: a knowledge-based approach*. San Mateo, CA: Morgan Kaufmann.
Rayner M, Carter D, Bouillon P *et al.* (2000). *The spoken language translator.* Cambridge, UK: Cambridge University Press.
Rozencvejg V J (ed.) (1974). *Machine translation and applied linguistics* (2 vols). Frankfurt: Athenaion Velag [Also published as: *Essays on lexical semantics* (2 vols). Stockholm: Skriptor.].
Slocum J (ed.) (1988). *Machine translation systems*. Cambridge, UK: Cambridge University Press.
Somers H L (2003). 'Machine translation: latest developments.' In Mitkov R (ed.) *The Oxford handbook of computational linguistics*. Oxford: Oxford University Press. 512–528.
TMI (1992). *Quatrième Colloque international sur les aspects théoriques et méthodologiques de la traduction automatique. Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation: Empiricist vs. rationalist methods in MT. Actes du colloque*. Montréal, Canada: CCRIT-CWARC.
Wahlster W (ed.) (2000). *Verbmobil: foundations of speech-to-speech translation*. Berlin: Springer.

## Relevant Websites

http://www.aamt.info – Asia-Pacific Association for Machine Translation, and its conferences.
http://www.aclweb.org – Association for Computational Linguistics and its conference and publication archive.
http://www.amtaweb.org – Association for Machine Translation, in the Americas, and its conferences.
http://www.eamt.org – European Association for Machine Translation, and its conferences.
http://www.elra.info – Conferences on Language Resources and Evaluation.
http://www.mt-archive.info – Machine Translation Archive.

# Machine Translation: Interlingual Methods

**B Dorr**, UMIACS, College Park, MD, USA
**E Hovy**, University of Southern California, Los Angelos, CA, USA
**L Levin**, Carnegie Mellon University, Pittsburgh, PA, USA





## Introduction

As described in the article on **Machine Translation: Overview**, machine translation (MT) methodologies are commonly categorized as direct, transfer, and interlingual. The methodologies differ in the depth of analysis of the source language and the extent to which they attempt to reach a language-independent representation of meaning or intent between the source and target languages. Interlingual MT typically involves the deepest analysis of the source language.

**Figure 1** – the Vauquois triangle (Vauquois, 1968) – illustrates these levels of analysis. Starting with the shallowest level at the bottom, direct transfer is made at the word level. Moving upward through syntactic and semantic transfer approaches, the translation occurs on representations of the source sentence structure and meaning, respectively. Finally, at the interlingual level, the notion of transfer is replaced with a single underlying representation – the interlingua – that represents both the source and target texts simultaneously. Moving up the triangle reduces the amount of work required to traverse the gap between languages, at the cost of increasing the required amount of analysis (to convert the source input into a suitable pretransfer representation) and synthesis (to convert the posttransfer representation into the final target surface form). For example, at the base of the triangle, languages can differ significantly in word order, requiring many permutations to achieve a good translation. However, a syntactic dependency structure expressing the source text may be converted more easily into a dependency structure for the target equivalent because the grammatical relations (subject, object, modifier) may be shared despite word order differences. Going further, a semantic representation (interlingua) for the source language may totally abstract away from the syntax of the language, so that it can be used as the basis for the target language sentence without change.

Comparing the effort required to move up and down the sides of the triangle to the effort to perform transfer, interlingual MT may be more desirable in