Research methods and system designs in machine translation: a ten-year review, 1984-1994

John Hutchins

University of East Anglia, UK.

1 The five eras of MT

Research on MT has passed through five eras to the present day. The first period began with the memorandum from Warren Weaver in 1949 which effectively launched MT research. The second began with the 1954 demonstration of a simple system for translation from Russian to English, which encouraged government agencies in the US and elsewhere to support large-scale projects. This period was brought to an end by the notorious ALPAC report in 1966, which highlighted the `failure' of MT research to meet its promises. The third, `quiet', era, when MT was virtually ignored, lasted until about 1975, with a revival of interest in Canada, Europe and Japan. Whereas the systems of the first two eras were generally based on the `direct' approach, the dominant framework after ALPAC was the various transfer and interlingual approaches based on linguistic rules. This fourth period lasted until the end of the 1980s with the emergence of corpus-based approaches (the use of bilingual text corpora, statistical methods, example-based approaches), and also the development of new rule-based methods using unification and constraint-based grammars. (For general descriptions of the historical development of MT see Hutchins 1986, 1988, 1993).

2 The situation in 1984

The revival of MT, which had begun some ten years after the ALPAC report of 1966, was firmly established by the mid 1980s. The SYSTRAN system had already been in use by the US Air Force for 15 years, for translating from Russian into English. The use of SYSTRAN at the Commission of the European Communities had started in 1977 with the initial English-French version, and since then a number of other language pairs had already been added. Since 1978 also SYSTRAN had been successfully operating at Xerox using a controlled English as input.

Elsewhere, in Canada, the sub language system Meteo for translating weather reports had been in operation since 1976; Fujitsu introduced its ATLAS systems in the early 1980s; at the same time, the first commercial MT systems appeared: ALPS and Weidner; and they were to be followed in the mid 1980s by a number of other, mainly Japanese commercial systems.

There were a few well-established research groups: the GETA team at Grenoble in France, the SUSY project at Saarbrücken in Germany, and the METAL project at the University of Texas (by now funded by the German Siemens Company). More significantly, the early and mid 1980s saw the establishment of a large number of MT research teams in Japan, primarily by the computer companies. There had been research in Japan in earlier years (notably at Kyoto University) but it came to international notice only in the 1980s. For Europe the most important development in MT research at this time was the establishment of the Eurotra project.

3 From 1984 to 1989

By the late 1970s it was widely assumed that systems such as SYSTRAN were based on approaches which were inherently incapable of improvement, and it was believed that the best prospects for improving the quality of MT output lay in the development of advanced 'indirect transfer' systems. It was also widely assumed that artificial intelligence (AI) could not, as yet, contribute much in the construction of large-scale MT systems.

The dominant framework of MT research from 1984 until the end of the 1980s was the approach based on essentially linguistic rules of various kinds: rules for syntactic analysis, lexical rules, rules for lexical transfer, rules for syntactic generation, rules for morphology, etc. Although the so-called `transfer' systems dominated, e.g. Ariane, Metal, SUSY, Mu and Eurotra, there appeared in the later 1980s various `interlingual' systems. Some were still essentially linguistics-oriented (DLT and Rosetta), but others adopted knowledge-based approaches, making use of non-linguistic information relevant to the domains of texts to be translated, in particular the research then beginning at Carnegie Mellon University, Pittsburgh. Nevertheless, these newer knowledge-based systems continued to be essentially rule-based systems, and in any case they remained somewhat of a novelty until almost the end of the decade.

3.1 Rule-based transfer systems

The Ariane, SUSY and Eurotra systems exemplified typical features of the so-called `second-generation' design: batch processing with post-editing and no interactive components, essentially syntax-oriented and stratificational with three stages of analysis, transfer and synthesis and with the processes of analysis and generation passing through series of distinct levels (morphology, syntax and semantics), with relatively abstract interface representations in the form of labelled trees, rules of transduction for changing trees from one level to another, and making little use of pragmatic and discourse information.

Throughout the 1980s the GETA team pioneered many innovations in MT design structure. In particular the Ariane system became the paradigm of the "second generation" linguistics-based `transfer' model. Its influence was profound, not only in Europe where the initial Eurotra concept owed much to the GETA research but also in Japan where the Mu system at Kyoto was based largely on Ariane, and which in its turn influenced many other Japanese researchers, and in South East Asia, particularly Malaysia and Thailand. Research at GETA covered an impressive variety of languages, but throughout it suffered from a lack of resources to develop substantial dictionaries, even in the language pairs most intensively investigated: Russian and French. Although Ariane as such did not become an operational system - there had been hopes in the mid 1980s that its Calliope project might result in a working system, but they came to nought - the system was thoroughly tested by B'VITAL and SITE, and in this decade it "lives on" in the Eurolang development (see section 4.2 below).

At Saarbrücken the experience was similar. During the 1980s the project team sought to demonstrate practical applications of its multilingual SUSY system. There was a collaborative project with the Kyoto University TITRAN system for translating document titles; there was the SUSANNAH project to develop a prototype translator's workstation; there was the establishment of STS, the Saarbrücker Translations-Service which provided translations for a number of German information centres. Research developments included the ASCOF project for French-German translation, which investigated newer parsing methods and the use of semantic networks for disambiguation, and in the SEMSYN project, a knowledge-based German text generation program which linked to the semantic interfaces produced by the Fujitsu ATLAS/II system. Like Ariane, SUSY itself did not become a working system, but it had a major influence on later systems.

The Eurotra project, initially planned in 1978 and formally instituted in 1982, was to be based on the latest advances in computational linguistics. It involved teams of researchers in each member nation of the Community. Its general design owed much to the GETA-Ariane and SUSY systems. Like them, it was a linguistics-based modular transfer system intended for multilingual translation producing good quality but not perfect output. The design combined lexical, logico-syntactic and semantic information in multilevel interfaces at a high degree of abstractness. No direct use of extra-linguistic knowledge bases or of AI-type inference mechanisms was made, and no facilities for human assistance or intervention during translation processes were to be incorporated. It assumed batch processing and human post-editing.

During the 1980s Eurotra stimulated innovative theoretical linguistic and computational-linguistic research, particularly in the Netherlands, Belgium, Denmark, Germany and Great Britain. Eurotra researchers advanced substantially the theoretical foundations of MT and made important contributions to syntactic theory (e.g. LFG and GPSG), formal parsing theory, and discourse analysis. One of the aims of the Eurotra project was to stimulate such research, and in this it succeeded. However, it did not produce a working prototype, and attempts towards the end of the project to involve industrial partnerships were largely unfruitful. Like Ariane, a major defect, readily conceded by those involved, was the failure to tackle problems of the lexicon, both theoretically and practically. While at the end of the 1970s, Eurotra was seen as representing the best `linguistics-based' design, at the end of the 1980s it was seen by many as basically obsolete in design and conception, and by 1992 Eurotra had effectively come to an end.

3.2 Rule-based interlingua systems

Such was the dominance of the 'transfer' approach in the early 1980s that few would have predicted the revival from the mid 1980s of the `interlingua' approach to MT. It was accepted that interlingual components would feature in advanced multilingual MT systems of the transfer type - and this in fact proved to be the case in the Eurotra research, where much attention was paid to `Euroversals' in transfer interfaces. But what would not have been widely anticipated would have been interlingua MT systems as such as the focus for research.

In the latter half of the 1980s there were two linguistics-based interlingua systems under development in the Netherlands: the DLT (Distributed Language Translation) project at the BSO company in Utrecht, and the Rosetta project at the Philips electronics company in Eindhoven. The six-year DLT project began in 1985 with support from a Netherlands ministry. It was intended as a multilingual interactive system operating over computer networks, where each terminal was to be a translating machine from and into one language only; texts were to be transmitted between terminals in an intermediary language. As its interlingua, DLT chose a modified form of Esperanto. Analysis was restricted primarily to morphological and syntactic features (formalised in a dependency grammar); there was no semantic analysis of the input. Disambiguation took place in the central interlingua component, where semantico-lexical knowledge was represented in an Esperanto database. From a combination of linguistic and extra-linguistic information the system computed probability scores for pairs of dependency-linked interlingual words. The project made a significant effort in the construction of large lexical databases, and in its final years proposed the building of a Bilingual Knowledge Bank from a corpus of (human) translated texts (Sadler 1989). In a number of respects DLT was a precursor of developments which have major importance in the mid 1990s (see sections 4.1 and 9 below).

The Rosetta project at Philips was innovative in another respect. The designers of this experimental system, involving three languages (English, Dutch and Spanish), opted to explore the use of Montague grammar in interlingual representations. A fundamental feature was the derivation of semantic representations from the syntactic structure of expressions, following the principle of compositionality; for each syntactic derivation tree there was to be a corresponding semantic derivation tree, and these semantic derivation trees were the interlingual representations.

A second feature was the exploration of the reversibility of grammars, and in this way Rosetta pioneered the direction of many current rule-based MT and NLP projects (see section 4.3 below).

A number of Japanese projects in the 1980s were also linguistics-based interlingua systems. Among them should be mentioned NEC's PIVOT system, which has been successfully demonstrated for English, Japanese, Korean, French and Spanish (Okumura et al. 1991).

The most important research on the rule-based interlingua approach, however, has been done at Carnegie Mellon University. The approach of this group combined linguistic analysis and semantico-conceptual knowledge bases. The work began originally at Colgate University in 1983, where a prototype system called TRANSLATOR was developed, until the main researchers moved to Carnegie Mellon, a major centre for research on artificial intelligence. From 1985 until 1989, the team worked on a knowledge-based MT system (KBMT89) (Goodman and Nirenburg (1991). The knowledge based approach is founded on the assumption that translation must go beyond linguistic knowledge and must involve `understanding' (for a general discussion of knowledge-based MT see Nirenburg et al. 1992.) Apart from familiar syntactic analysis and generation components, the KBMT89 model includes a `mapping rule interpreter' for converting LFG-type structures into semantically interpreted representations and an interactive `augmentor' for residual ambiguities. The semantic mapper draws information from a knowledge database of the domain (initially computer manuals). Interlingual representations are intended to convey the `actual events' of source texts as networks of fully interpreted (expanded) propositions, i.e. events or states with their arguments and causal, temporal, spatial, etc. links to other events or states. An important feature is the representation of anaphoric links, textual relations, speech acts, topic-comment relationships, etc. A number of interlingua models have been developed, notably KANT and CATALYST (see 4.2 below), and the team has devoted much work on problems of knowledge and lexical acquisition (see 5 below).

4 From 1989 to the present

4.1 Corpus-based methods

To the end of the 1980s the rule-based approach dominated. Since 1989 this dominance has been broken by the emergence of new methods and strategies which are now loosely called `corpus-based' methods. Firstly, a group from IBM published in 1988 the results of experiments on a system based purely on statistical methods. The effectiveness of the method was a considerable surprise to many researchers and has inspired others to experiment with statistical methods of various kinds in subsequent years. Secondly, at the very same time certain Japanese groups began to publish preliminary results using methods based on corpora of translation examples, i.e. using the approach now generally called `example-based' translation. For both approaches the principal feature is that no syntactic or semantic rules are used in the analysis of texts or in the selection of lexical equivalents.

The most dramatic development was the revival of the statistics-based approach to MT in the Candide project at IBM. Statistical methods were common in the earliest period of MT research, in the 1960s, but the results had been disappointing. With the success of newer stochastic techniques in speech recognition, the IBM team at Yorktown Heights began to look again at their application to MT. The distinctive feature of Candide is that statistical methods are used as virtually the sole means of analysis and generation; no linguistic rules are applied. The IBM research (Brown et al. 1990) is based on the vast corpus of French and English texts contained in the reports of Canadian parliamentary debates (the Canadian Hansard). The essence of the method is first to align phrases, word groups and individual words of the parallel texts, and then to calculate the probabilities that any one word in a sentence of one language corresponds to a word or words in the translated sentence with which it is aligned in the other language.

What surprised most researchers (particularly those involved in rule-based approaches) was that the results were so acceptable: almost half the phrases translated either matched exactly the translations in the corpus, or expressed the same sense in slightly different words, or offered other equally legitimate translations. Obviously, the researchers have sought to improve these results, and the IBM group proposes to introduce more sophisticated statistical methods (Brown et al. 1992). But they also, rather surprisingly, intend to make use of some minimal linguistic information. Although they set out to disprove the traditional linguistic rule-based approaches, they are ready to experiment with any method which gives good results! Among the proposals are the treatment of all morphological variants of a verb as a single word, and the use of syntactic transformations to bring source structures closer to those of the target language.

The second major `corpus-based' approach - benefiting likewise from improved rapid access to large data banks of text corpora - is what is known as the `example-based' (or `memory-based') approach. Although first proposed in 1984 (Nagao 1984), it was only towards the end of the 1980s that experiments began, initially in some Japanese groups and during the DLT project (as already mentioned). The underlying hypothesis is that translation often involves the finding or recalling of analogous examples, the discovery or recollection of how a particular expression or some similar phrase has been translated before. The example-based approach is founded on processes of extracting and selecting equivalent phrases or word groups from a data bank of parallel bilingual texts, which have been aligned either by statistical methods (similar perhaps to those used by the IBM group) or by more traditional `rule-based' morphological and syntactic methods of analysis. For calculating matches, some MT groups use semantic methods, e.g. a semantic network or a hierarchy (thesaurus) of domain terms. Other groups use statistical information about lexical frequencies in the target language. The main advantage of the approach is that since the texts have been extracted from data banks of actual translations produced by professional translators there is an assurance that the results will be accurate and idiomatic.

At present, the example-based approach has been used most often to complement more traditional methods based on linguistic rules, e.g. at ATR (Furuse & Iida 1992) and in IBM Japan's SHALT system (Takeda et al. 1992). However, there are some researchers who contend that the effectiveness and general validity of the approach can be fully tested only if it is used as the sole method of generating target text.

The availability of large corpora has encouraged experimentation in methods deriving from the computational modelling of cognition and perception, in particular research on parallel computation, neural networks or connectionism. In natural language processing, connectionist models are `trained' to recognise the strongest links between grammatical categories (in syntactic patterns) and between lexical items (in semantic networks).

The potential relevance to MT is clear enough for both analysis and transfer operations, given the difficulties of formulating accurate grammatical and semantic rules in traditional approaches. As yet, however, within MT only a few groups have done some small-scale research in this framework, e.g. in the speech translation research at Carnegie Mellon University (Jain et al. 1991), in an example-based approach by McLean (1992) at UMIST, and in the Matsushita transfer-based prototype system (Ishikawa & Sugimura 1992).

Connectionism offers the prospect of systems `learning' from past successes and failures. So far the nearest approach has been for systems to suggest changes, on the basis of statistics, about corrections made by users, e.g. during post-editing. This model is seen in the commercial Tovna system and in the experimental PECOF `feedback' mechanism in the Japanese MAPTRAN system (Nishida & Takematsu 1990). A similar mechanism has been incorporated in the NEC PIVOT system (Miura et al. 1992).

4.2 Rule-based MT since 1990

Although the main innovation of recent years has been the growth of corpus-based approaches, rule-based research continues. To a certain extent, both the Ariane and Eurotra projects have survived in so far as the Eurolang project is basing its MT research on their foundations. This is based at SITE, a French company which purchased B'VITAL, the Grenoble company founded to develop ARIANE in the French National MT project Calliope. Also involved is the German company Siemens-Nixdorf which is to contribute with its METAL system. Initially Eurolang will develop ten language pairs, English into and from French, German, Italian and Spanish, and French into and from German. The project is to be based on Ariane and METAL and will build upon the expertise of research teams involved in the Eurotra project. However, the first product of Eurolang has not been an MT system as such but a translator's workstation, the Optimiser, which eventually will incorporate an MT module.

Although the Eurotra project itself did not produce a working system, a number of researchers involved continued to work on the theoretical approach developed, e.g. the CAT2 system at Saarbrücken (Sharp & Streiter 1992). One of the fruits of Eurotra research has been the PaTrans transfer-based system developed in Denmark for Danish/English translation of patents (Hansen 1994). Elsewhere, research on linguistics-based transfer systems continues in the present decade. Research proceeds on the Siemens' system METAL, which reached the market at the end of the 1980s with a German-English version, and on which development continues at various European locations on further language pairs. There is also the LMT project which began in the mid-1980s, based at a number of IBM research centres in Germany, Spain, Israel and the United States. Translation is via four steps implemented in Prolog: lexical analysis, producing descriptions of input words and their transfers; syntactic analysis of source texts, producing representations of both surface and deep (logical) relations; transfer, involving both isomorphic structural transfer and restructuring transformations; and morphological generation of target texts. Slot grammar is characterised as combining a lexicalist approach to grammar and logic programming - LMT stands for `Logic programming MT'. The language pairs under investigation include English-German, German-English, and English-Spanish (McCord 1989; Rimon et al. 1991).

While the DLT and Rosetta projects have ended, major `interlingual' projects continue to thrive, indeed with even more vigour, particularly in the knowledge-based approach at Carnegie Mellon University. Several models have been developed over the years, and in 1992 was announced the beginning of a collaborative project with the Caterpillar company with the aim of creating a large-scale high-quality system CATALYST for multilingual translation of technical manuals in the specific domain of heavy earth-moving equipment.

Other interlingua-based systems are e.g. the ULTRA system at the New Mexico State University (Farwell & Wilks 1991) and the UNITRAN system based on the linguistic theory of Principles and Parameters (Dorr 1993). There is also the Pangloss project (Frederking 1994), an interlingual system restricted to the vocabulary of mergers and acquisitions, a collaborative project involving experts from the universities of Southern California, New Mexico State and Carnegie Mellon. Pangloss is itself one of three MT projects supported by ARPA, the others being the IBM statistics-based project (see above) and a system being developed by Dragon Systems, a company which has been particularly successful in speech research but with no previous experience in MT.

4.3 Lexicalist tendency, and constraint-based formalisms

A characteristic feature of rule-based systems is the transformation or mapping of labelled tree representations. For example, in Eurotra a series of tree transductions was proposed: from a morphological tree into a syntactic tree, from a syntactic tree into a semantic tree, from an interface tree of the source language into an equivalent target-language tree, and so forth. Transduction rules require the satisfaction of precise conditions: a tree must have a specific structure and contain particular lexical items or specific syntactic or semantic features. In addition, every tree is tested by formation rules; in effect, a `grammar' confirms the acceptability of its structure and the relationships it represents. A tree is rejected if it does not conform to the grammatical rules of the level in question: morphological, syntactic, semantic, etc. Grammars and transduction rules specify the `constraints' which determine the possibility of transfer from one level to another and hence, in the end, the transfer of a source-language text to a target-language text.

Since the mid 1980s there has emerged a widely accepted general framework for rule-based systems. It embraces all the formalisms which can be categorised as variants or equivalents of `unification' and `constraint-based' formalisms. In essence, what these formalisms have in common is that the large set of rules devised only for application in very specific circumstances and to specific representations has been replaced by a restricted set of abstract rules and the incorporation of the conditions and constraints into specific lexical entries. The transformation rules themselves are now expressed as operations of rules of unification, which control the interaction of sets of features, the formation of new sets and the elimination of illegitimate sets. As a result, the syntactic orientation which characterised many transfer systems in the past has been replaced by a trend towards lexicalist solutions. Many current research projects illustrate the tendency, including the LMT and UNITRAN system already mentioned.

An extreme example of the `lexicalist' approach is the method known as "shake and bake" (Whitelock 1992). There are no longer any structural representations, there are only sets of lexical representations. Translation proceeds through the identification of lexical items in the target language which satisfy the semantic constraints which have been attached to the equivalent lexical items in the source language. A translation is produced (or `baked') from interactions among the sets of features and the constraints attached to target language words.

Unification grammar and constraint-based grammars originated some ten years ago (Kay 1984). Since the late 1980s, unification has become a central concept for a large number of linguistic theories, and constraint-based grammars and formalisms dominate rule-based MT research: e.g. Lexical Functional Grammar, Definite Clause Grammar, Head-driven Phrase Structure Grammar, Categorial Grammar, etc. The main advantage of these grammars is the simplification of the rules (and hence the computational processes) of analysis, transformation and generation. Instead of a series of complex multi-level representations there are mono-stratal representations or simple lexical transfer. At the same time, the components of these grammars are in principle reversible. It is no longer necessary to construct for the same language different grammars of analysis and generation: the same formalism and the same grammars can in theory be applied in both directions.

Several general-purpose NLP systems based on unification and constraint-based grammars have been applied to translation tasks. The CLE (Core Language Engine) system, for example, has been used for automatic translation from Swedish into English and vice versa (Alshawi 1992); the PLNLP (Programming Language for Natural Language Processing) system provided the foundation for translation systems involving English, Portuguese, Chinese, Korean and Japanese (Jensen 1993); and the ELU engine (Environnement Linguistique d'Unification) developed at

Geneva in Switzerland has formed the basis for a bi-directional system for translating avalanche bulletins between French and German (Bouillon & Boesefeld 1992). The testing of general-purpose NLP systems on translation is likely to increase, both because MT represents a good testbed for evaluation and because MT products have an obvious market potential.

5 Lexicon acquisition

The trend towards lexicalist approaches has had an important impact on the construction of lexicons. With the increase in the range of information attached to lexical units the lexicon is no longer concerned just with morphological and grammatical data of source language words and with indicating equivalent words or phrases in target languages. It includes now information on syntactic and semantic constraints and non-linguistic and conceptual information, albeit often limited to restricted subject domains. The expansion of data has been most clearly seen in the lexicons of interlingua-based systems which include large amounts of non-linguistic information. Hence the high priority of this research task for the Carnegie Mellon, LMT and UNITRAN teams (see Gates & Shell 1993, Neff & McCord 1990, Dorr & Voss 1993)

In recent years interest has grown rapidly in addressing the problems of constructing lexicons for MT, and a number of workshops devoted to the question have been held (AAAI 1993). Lexicon building is a complex and expensive task if the lexicon is to be adequate and sufficient for real and practical applications in operational situations. Many MT research groups are investigating methods of acquiring lexical information from readily available lexicographic sources, such as bilingual dictionaries intended for language learners, specialised technical dictionaries, and the terminological data banks used by professional translators (e.g. EDR 1990). At the same time, research groups are collaborating more closely with each other in projects for the construction of lexicons for a wide range of natural language applications and different types of systems, not just for machine translation but also for text analysis and information retrieval. The best known collaborative project in the MT field is the EDR project (Electronic Dictionary Research) supported by several Japanese computer manufacturing companies (Takebayashi 1993)

6 Generation and database MT

The example-based approach has strengthened a trend which was already evident in the `rule-based' framework, namely the much greater attention paid to questions of generating good quality texts in target languages. Ten years ago in the mid 1980s it was commonly believed that the most difficult problems of MT concerned syntactic and semantic analysis, the disambiguation of homonyms, the resolution of structural ambiguity, and the identification of the antecedents of pronouns; in other words, the main problem area of MT was the understanding of the text to be translated. The thrust of research on linguistic rules and on knowledge bases reflected this concentration on problems of analysis. At that time, the problem of generating idiomatic output text in the target language was a largely neglected area of MT research. Now, major efforts are devoted to questions of stylistic improvement of output and to discourse features (e.g. DiMarco 1994, Mitkov 1992)

Much of the impetus for this research has come from increasing attention to the need to provide natural language output from searches in databases. The `language' of the database content is necessarily artificial and constrained in various ways; providing comprehensible output can be regarded as a mode of MT from an artificial language to a natural language. While most of this research concentrates on generating text in a single language, some of it is devoted to multilingual generation. One of the first group to tackle this topic was a team of researchers based in Montreal long involved in MT. This group has worked on a system for producing marine forecasts in French and English (Bourbeau et al. 1990), and on a system for generating bilingual summaries of statistical data on the labour force (Iordanskja et al. 1990). Similar research on generating weather forecasts is reported by Mitkov (1992).

Another important trend of the last five years is the recognition of a demand for types of translations which have not previously been studied. In the past, systems were built generally for bilingual users, for translators and for those knowing both source and target languages. In addition, the texts translated had to be post-edited. The needs of those wanting to translate into languages they do not know were neglected. Businessmen engaged in foreign trade often need to communicate fairly simple standard messages in an unknown language (e.g. confirmation of an order, booking of accommodation, etc.) In recent years, groups have experimented with `dialogue-based MT' systems where the text to be translated is composed or written in a collaborative process between man and machine (e.g. at UMIST, the University of Brussels, Grenoble University and at the Science University of Malaysia; Somers 1992, Jacqmin 1992, Boitet 1990, Zaki et al. 1991). In this way it is possible to construct a text which the system is known to be capable of translating without further reference to the author, which needs no revision and for which good quality output can be assured. This type of MT might be called `MT for writers'.

7 Controlled language, domain-specific, sub language, and custom-built MT

In practice nearly all MT systems have been largely limited to restricted domains. Although originally designed as general-purpose systems, many of the well-established systems have been limited in operation to particular ranges of subjects, since large dictionaries are needed and developers have concentrated on domains where there is greatest demand. This trend was already well established by the mid 1980s. Many systems have been specifically designed for particular subject areas (`sub languages') or for the needs of specific users. In each case, there are efforts to overcome the known deficiencies of full-scale MT, in particular the difficulties of analysing complex sentences, of selecting correct target language equivalents and of generating idiomatic output. Consequently, the same systems may feature combinations of the three options: (a) control of input texts, (b) restriction to a sub language, and (c) design for a specific user.

The control of the vocabulary and of the grammatical structures of texts submitted for translation reduces the difficulties of constructing satisfactory lexicons of sufficient coverage, and the problems of ambiguity and selection of equivalents. Although the costs of preliminary editing may be high, post-editing is reduced considerably. The Xerox implementation of SYSTRAN and the many successful systems developed by the Smart Corporation were well known examples of controlled language MT during the 1980s. The trend continues in the 1990s, e.g. one of the largest controlled language projects currently is the CATALYST system under development for Caterpillar Corporation. Whereas controlled language has previously been used in systems of the `direct translation' design, this will be the first application in a more advanced `interlingua' system.

The design of systems for a specific sub language is also not new: the well known Meteo has been translating meteorological reports for 15 years. Among the sub language systems of recent years there are the CRITTER system for reports on the stock market under development in Montreal (Isabelle et al. 1988), the already mentioned projects at ELU, Pangloss, and the extremely ambitious projects for the development of spoken language translation (see section 8 below.)

In the past, there were few systems built by users themselves. Exceptions in the early 1980s were the systems built at PAHO (Pan American Health Organisation), for translating from English into Spanish (ENGSPAN) and from Spanish into English (SPANAM). Since 1990 there has been a considerable increase in user-designed systems - typically with restricted vocabularies, for a particular domain and often based on a specific sub language. Some of these systems have been developed by software companies for clients. For example, Volmac Lingware Services has produced MT systems for a textile company, an insurance company, and for translating aircraft

maintenance manuals (Van der Steen & Dijenborgh 1992); Cap Gemini Innovation developed TRADEX to translate military telex messages for the French Army (Aumaitre et al. 1992); and in Japan, CSK developed its own ARGO system for translation in the area of finance and economics, and now offers it also to outside clients (Carbonell et al. 1992). Such user-designed systems are a sign that the computational methods of MT and NLP are now becoming familiar outside the limited circles of researchers. The systems may perhaps only rarely be innovative from a theoretical or methodological point of view, but they are often very advanced computationally. It is a trend which is expected to expand rapidly in coming years.

8 Speech translation

One of the most significant developments of the last five years has been the growing interest in spoken language translation. Pioneer work was done at British Telecom and at the ATR laboratories near Osaka, Japan. The latter has been a major well funded project supported by government and industry (Morimoto & Kurematsu 1993). The first period of the project extended from 1986 to 1993; the second period will last until the end of the century. The ATR team is developing a system for telephone registrations at international conferences and for telephone booking hotel accommodation. The project combines an impressive range of basic research in speech recognition and linguistic analysis of conversational language.

In 1992, ATR joined in a collaborative project with Carnegie Mellon University and Karlsruhe University. Carnegie Mellon had experimented with a spoken language translation in its JANUS project during the late 1980s. Since the early 1990s the University of Karlsruhe has been involved in an expansion of the JANUS system. The three groups have formed a consortium C-STAR (Consortium for Speech Translation Advanced Research), each developing speech recognition and speech synthesis modules for their own languages (Japanese, English, German) and translation programs linking their language to the other two (Woszczyna et al. 1993). In January 1993 the consortium gave a successful public demonstration of telephone translation from English, German and Japanese each into the other two languages. The system is currently limited in its domain to conference registrations. Recently a second phase has begun involving further groups in France (LMSI), the United Kingdom (SRI) and other groups in North America and Asia.

More recently still, in May 1993, there began the German funded Verbmobil project (Wahlster 1993). This is an 8-10 year project aiming to develop a transportable aid for face to face English-language commercial negotiations by Germans and Japanese who do not know English fluently. The goal is a research model in 1997 and, four years later, a prototype product. The Verbmobil project will also be collaborating with the ATR Laboratories.

9 Third generation systems

This paper has illustrated marked differences between the types of systems developed in the mid 1980s and the kind of research being conducted in the 1990s. There has been a change of emphasis in many aspects. The dominance of rule-based approaches has been broken by the appearance of a variety of corpus-based methods: statistical MT, example-based MT, connectionist approaches, research on spoken translation, etc. There has been the development of new types of MT systems for monolingual users, for specific domains and with controlled vocabularies. Within the rule-based systems there has been a move away from the syntax orientation of the past to a `lexicalist' position, from tree transduction rules to constraint-based methods, from an emphasis on analysis, disambiguation and `understanding' to the generation of fluent stylistically appropriate quality output, from the compilation of linguistic information for lexical entries to the acquisition of lexical and conceptual knowledge bases from existing dictionaries and text corpora. In sum, the "second generation" system type of the 1980s - i.e. the linguistics rule-based `transfer' model - is now seen as outmoded. However, it is not clear yet

whether the new era which began in 1990 will produce its own "third generation" paradigm. Most observers agree that future research systems (and eventually commercial systems) will be `hybrids' of linguistics rule-based, statistics-based and example-based methods. But the details are uncertain.

In one possible perspective, the well-proven linguistic methods of the `indirect' systems will provide the foundation upon which processes involving domain-specific knowledge banks, statistical data and examples of translated texts will operate. The strength of the rule-based approaches lies in the reliability of syntactic and morphological analyses. The potential strengths of the newer methods lie in domain-oriented semantic analysis where linguistic rules have been insufficient for reliable disambiguation and for producing good quality output.

In the "third generation" systems, therefore, we may find that syntactic analysis is limited to the recognition of surface structures, phrase constituents and dependency relations, and there will be almost no deep analysis of logical relations (quantification, scope of negation). Semantic analysis may be limited to the identification of `grammatical' roles: agent, instrument, etc. For these purposes relatively sparse lexical information may be all that is required and could be easily extracted from standard sources such as general-purpose dictionaries. As a consequence, the rules for lexical and structural transfer may apply to much shallower representations than we are familiar with in systems such as Ariane and Eurotra. Also the rule-based formalisms are likely to be the unification and constraint-based grammars which are now becoming standard frameworks for much NLP research.

The corpus-based methods will provide the means for refining and enhancing the lexical and semantic aspects of analysis, transfer and generation, and will thus replace the intractable complexity of rule-based approaches. Widespread use can be envisaged for access to translation examples stored in aligned bilingual text banks; this is already becoming an increasingly common feature of translation workstations, and will undoubtedly feature in most future MT systems. There will be increasing use of statistical information on lexical collocations and monolingual vocabulary frequencies for assisting disambiguation during syntactic and semantic analysis of phrases, for selecting appropriate idiomatic target language phrases. These statistical methods are likely to be tied closely to domain-specific knowledge banks and terminological data banks in these disambiguation and selection tasks. Finally, we may expect more attention to feedback and/or connectionist methods to improve grammars (and/or rule bases) and to enhance monolingual and bilingual lexicons.

10 Wider perspectives

In describing developments in the design and methodology of MT systems over the past ten years, most examples have been taken from research activity in Western Europe, North America and Japan. A distinctive feature of the last decade, however, has been the globalisation of MT research. Within the last five years, research activity has grown rapidly in China, Taiwan, Korea, India and South East Asia. Major projects have been initiated involving international co-operation - the CICC multilingual project is just one example (Funaki 1993). By contrast, MT research in Eastern Europe has been affected profoundly by the political changes since 1989. In many cases, researchers have successfully joined collaborative projects with Western European groups, e.g. researchers from the Czech Republic and from Bulgaria; others have been able to continue only at a much reduced level, particularly researchers from the former Soviet Union. As this paper has shown, MT research in Western Europe is in a transitional stage; the sponsorship of the European Commission has diminished and industrial support has increased with a move away from large-scale projects to company-based shorter-term developments. In North America during the 1980s, MT research was still only recovering gradually from the effects of the ALPAC report; in the 1990s, United States research is thriving and has recaptured its former theoretical innovativeness.

This paper has concentrated on changes in the design and development of what may be called the `MT engine', the central component of systems. But these are only part of much wider and perhaps more profound changes. Recent years, particularly since 1990, have demonstrated the crucial and central importance of MT as part of the overall `industrialisation' of all aspects and all forms and types of language and information work, whether monolingual or multilingual, whether written texts or spoken language, whether based on `traditional' documents or involving semi-composed texts, whether produced in printed form or transmitted over computer networks.

During the last decade it has become apparent that the traditional conception of MT as a process which runs parallel to that of the professional translator (and which thus threatens to replace him or her) is too narrow. The greatest expansion in the use of MT systems has taken place in commercial agencies, government services and multinational companies providing translations on a large scale, primarily technical documentation (Vasconcellos 1993), and this growth is likely to increase even faster in the coming years. However, there has also been a rapid increase in the use of MT for occasional translation, both through the provision of on-line services such as Minitel and a number of Japanese networks, and from the availability on the market of cheap PC-based software such as Globalink, PC-Translator and MicroTac.

These uses do not by any means exhaust the potentiality of MT. From the beginning, MT systems have been used to provide rapid, unedited versions of scientific and technical documentation for experts familiar with the topic and not concerned about the translation quality. This need still continues, but few if any systems are designed specifically for this type of use. One focus of future MT research could be the development of `MT systems for surveillance' which might provide rapid low-quality translation of the gists or abstracts of documents from a wide range of languages. A related need is the provision of translated textbooks for developing countries, generally from the `major' Western languages into languages of Asia and Africa which have so far been neglected by MT research. Another area for MT research, which has however already been recognised, is the development of systems for monolingual users who want to translate from their own language into another partially or wholly unknown language. The primary need is the translation of relatively standardised messages into corresponding and appropriate forms of the target languages, with the assurance of good quality output. Already recognised also is the potentially huge market for speech translation, initially within limited domains, and research in this area was mentioned above. Finally, the growth of telecommunication networks and access to information sources and databases opens a further vast opportunity for research on new types of MT systems for the on-line translation of electronic mail, bulletin boards, discussion lists and other textual information (a beginning has already been made by CompuServe); for providing multilingual access to databases and the translation on-line of the abstracts and texts retrieved; and so forth.

What is absent from this list, it will be noted, is any mention of `MT for translators'. It is quite clear from recent developments that what professional translators need are tools to assist them to translate: access to dictionaries and terminological data banks, multilingual word processing, management of glossaries and terminology resources, input and output communication (e.g. OCR scanners, electronic transmission, high-class printing). For these reasons, the most appropriate and successful developments of the last few years have been the translator workstations. One of the most significant additions to their range of facilities has been the `translation memory' which enables the storage of and access to existing translations for later (partial) reuse or revision or as sources of example translations. In this development the research on bilingual text alignment mentioned above (section 4.1) is of central importance.

Translator workstations represent one example of the growing trend in computer-based document processing. In the future, much MT research will be oriented towards the development of `translation modules' to be integrated in such general `office' systems, rather than the design of systems to be self-contained and independent. It is already evident that the range of computer-based `translation' activities is expanding to embrace any process which results in the production

or generation of texts and documents in bilingual and multilingual contexts, and it is quite possible that MT will be seen as the most significant component in the facilitation of international communication and understanding in the future `information age'.

References

The sources for this paper are mainly the proceedings of the biennial MT Summit conferences (Hakone 1987, Munich 1989, Washington 1991, Kobe 1993) and of the international conferences on Theoretical and Methodological Issues in Machine Translation (TMI) at Austin 1990, Montreal 1992, and Kyoto 1993. For details of systems before 1990 see references in Hutchins 1986 and 1988.

AAAI (1993): Building lexicons for machine translation: Papers from the 1993 AAAI Spring Symposium, March 23-25, Stanford, California. Menlo Park, Ca.: AAAI Press.

Alshawi, H. (1992), ed. *The Core Language Engine*. Cambridge, Mass.: MIT Press.

AMTA (1994): Technology Partnerships for Crossing the Language Barrier. Proceedings of the First Conference of the Association for Machine Translation in the Americas, 5-8 October 1994, Columbia, Maryland.

Aumaitre, J. M. et al. (1992): TRADEX, un système de traduction de telex. *Meta* 37 (4), 624-634.

Bouillon, P. & Boesefeldt, K. (1992): Problèmes de traduction automatique dans le sous-langage des bulletins d'avalanches. *Meta* 37(4), 635-646.

Boitet, C.(1990): Towards personal MT: general design, dialogue structure, potential role of speech. In: Coling 90 (3), 30-35.

Bourbeau, L. et al. (1990): Bilingual generation of weather forecasts in an operations environment. In: Coling 90 (1), 90-92.

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

Brown, P.F. et al. (1992): Analysis, statistical transfer, and synthesis in machine translation. In: TMI-92, 63-100.

Carbonell, J. et al. (1992): *JTEC Panel report on machine translation in Japan*. Baltimore, MD: Japanese Technology Evaluation Centre.

Coling 88: *Coling Budapest: Proceedings of the 12th International Conference on Computational Linguistics*. Budapest: John von Neumann Society for Computing Sciences, 1988.

Coling 90: *Coling-90: papers presented to the 13th International Conference on Computational Linguistics*...1990 Aug, Helsinki, Finland. Helsinki: Yliopistopaino, 1990. 3 vols.

Coling 92: *COLING-92: proceedings of the fiftieth International Conference on Computational Linguistics*, 23-28/8/1992, Nantes. Grenoble: GETA, 1992. 4 vols.

DiMarco, C. (1994): Stylistic choice in machine translation. In: AMTA (1994), 32-39.

Dorr, B.J. (1993): *Machine translation: a view from the lexicon*. Cambridge, Mass.: MIT Press, 1993.

Dorr, B.J. & Voss, C.R. (1993): Constraints on the space of MT divergences. In: AAAI (1993), 43-53.

EDR (1990): Proceedings of International Workshop on Electronic Dictionaries, November 8-9, 1990, Oiso, Kanagawa, Japan. Tokyo: EDR.

Farwell, D. & Wilks, Y. (1991): ULTRA: a multilingual machine translator. In: MT Summit 3, 19-24.

Frederking, R. et al. (1994): Integrating translations from multiple sources within the Pangloss Mark III machine translation system. In: AMTA (1994), 73-80.

Funaki, S. (1993): Multi-lingual Machine Translation (MMT) project. In: MT Summit 4, 73-78.

Furuse, O. & Iida, H. (1992): Transfer-driven machine translation. In: International Workshop on Fundamental Research for the Future Generation of Natural Language Processing, 30-31 July 1992, Centre for Computational Linguistics, UMIST; 95-111.

Gates, D.M. & Shell, P. (1993): Rule-based acquisition and maintenance of lexical and semantic knowledge. In: Sixth Conference of the European Chapter of the Association for Computational Linguistics, 21-23 April 1993, Utrecht University; 149-157.

Goodman, K. & Nirenburg, S. (1991), eds.: *The KBMT project: a case study in knowledge-based machine translation*. San Mateo, Ca.: Morgan Kaufmann.

Hansen, V. (1994): PaTrans - a MT system: development and implementation of and experiences from a MT-system. In: AMTA (1994), 114-121.

Hutchins, J. (1986): *Machine translation: past, present, future*. Chichester: Ellis Horwood.

Hutchins, J. (1988): Recent developments in machine translation. In: Maxwell, D. et al. (eds.) New directions in machine translation. Conference proceedings, Budapest 18-19 August 1988. Dordrecht: Foris; 7-63.

Hutchins, J. (1993): Latest developments in machine translation technology. In: MT Summit 4, 11-34.

Iordanskaja, L. et al. (1992): Generation of extended bilingual statistical reports. In: Coling 92 (3), 1019-1023.

Isabelle, P. et al. (1988): CRITTER: a translation system for agricultural market reports. In: Coling 88, 261-266.

Ishikawa, M. & Sugimura, R. (1992): Natural language analysis using a network model: modification deciding network. In: TMI-92, 55-66.

Jacqmin, L.(1992). La traduction automatique au service de l'utilisateur monolingue. *Meta* 37(4), 610-623.

Jensen, K. et al. (1993): *Natural language processing: the PLNLP approach*. Boston: Kluwer.

Kay, M. (1984): Functional unification grammar: a formalism for machine translation. In: *Coling 84: 10th International Conference on Computational Linguistics*, 2-6 July 1984, Stanford University, California; 75-78.

McCord, M. C. (1989): Design of LMT: a Prolog-based machine translation system. *Computational Linguistics* 15, 33-52.

McLean, I.J. (1992): Example-based machine translation using connectionist matching. In: TMI-92, 35-43.

Mitkov, R. (1991): Generating public weather reports. In: Proceedings International Conference on Current Issues in Computational Linguistics, 12-14 June 1991, Universiti Sains Malaysia, Penang, Malaysia; 372-378.

Mitkov, R. (1992): Discourse-based approach in machine translation. In: International Symposium on Natural Language Understanding and AI, July 12-15, 1992, Kyushu Institute of Technology, Iizuku, Fukuoka; 225-230.

Miura, M. et al. (1992): Learning mechanism in machine translation system "PIVOT". In: Coling 92, (2), 693-699.

Morimoto, T. & Kurematsu, A. (1993): Automatic speech translation at ATR. In: MT Summit 4, 83-96

MT Summit 3: *MT Summit III, July 1-4 1991*, Washington D.C., USA.

MT Summit 4: *MT Summit IV: International Co-operation for Global Communication, July 20-22, 1993*, Kobe, Japan.

Nagao, M. (1984): A framework of a mechanical translation between Japanese and English by analogy principle. In: Elithorn, A. & Banerji,R.(eds.) *Artificial and human intelligence*. Amsterdam: North-Holland; 173-180.

Neff, M.S. & McCord, M.C. (1990): Acquiring lexical data from machine-readable dictionary resources for machine translation. In: TMI-90, 85-90.

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

Nishida, F. & Takematsu, S. (1990): Automated procedures for the improvement of a machine translation system by feedback from postediting. *Machine Translation* 5(3), 223-246.

Okumura, A. et al. (1992): A pattern-learning based hybrid model for the syntactic analysis of structural relationships among Japanese clauses. In: TMI-92, 45-54.

Rimon, M. et al.(1991): Advances in machine translation research in IBM. In: MT Summit 3, 11-18.

Sadler, V. (1989): *Working with analogical semantics: disambiguation techniques in DLT*. Dordrecht: Foris.

Sharp, R. & Streiter, O. (1992): Simplifying the complexity of machine translation. *Meta* 37 (4), 681-692.

Somers, H.L. (1992): Interactive multilingual text generation for a monolingual user. In: TMI-92, 151-161.

Takebayashi, Y. (1993): EDR electronic dictionary. In: MT Summit 4, 117-126

Takeda, K. et al. (1992): Shalt2 - a symmetric machine translation system with conceptual transfer. In: Coling 92 (3), 1034-1038.

TMI-90: *Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, 11-13 June 1990, Austin, TX, USA.

TMI-92: *Fourth International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages. Empricist vs Rational Methods in MT*, June 25-27, 1992, Montreal, Canada.

TMI-93: *Fifth International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages. MT in the Next Generation*, July 14-16, 1993, Kyoto, Japan.

Van der Steen, G. & Dijenborgh, B.J.(1992): On-line correction and translation of industrial texts. In: *Translating and the computer 14* (London: ASLIB), 135-164.

Vasconcellos, M. (1993): The present state of machine translation usage technology, or: How do I use thee? Let me count the ways! In: MT Summit 4, 35-46. Also in: *MT News International* 6 (September 1993), 12-17.

Wahlster, W. (1993): Verbmobil: translation of face-to-face dialogs. In: MT Summit 4, 127-135

Whitelock, P. (1992): Shake-and-bake translation. In: Coling 92 (2), 784-791.

Woszczyna, M. et al. (1993): Recent advances in JANUS: a speech translation system. In: TMI-93, 195-200.

Zaki Abu Bakar, A. et al.(1991): Malay official letters translation. In: Proceedings International Conference on Current Issues in Computational Linguistics, 12-14 June 1991, Universiti Sains Malaysia, Penang, Malaysia; 413-420.