[*Machine Translation Today: Translating and the Computer 15*. Papers at a conference…18-19 November 1993 (London: Aslib)]

A NEW ERA IN MACHINE TRANSLATION RESEARCH

John Hutchins
University of East Anglia, Norwich, England

In the 1980s the dominant framework of MT was essentially 'rule-based', e.g. the linguistics-based approaches of Ariane, METAL, Eurotra, etc.; or the knowledge-based approaches at Carnegie Mellon University and elsewhere. New approaches of the 1990s are based on large text corpora, the alignment of bilingual texts, the use of statistical methods and the use of parallel corpora for 'example-based' translation. The problems of building large monolingual and bilingual lexical databases and of generating good quality output have come to the fore. In the past most systems were intended to be general-purpose; now most are designed for specialised applications, e.g. restricted to controlled languages, to a sublanguage or to a specific domain, to a particular organisation or to a particular user-type. In addition, the field is widening with research under way on speech translation, on systems for monolingual users not knowing target languages, on systems for multilingual generation directly from structured databases, and in general for uses other than those traditionally associated with translation services.

INTRODUCTION

At the end of the 1980s, machine translation entered a period of innovation in methodology which has changed the framework of research.

What has changed? What was the situation in MT five years ago? Between 1975 and 1988 a large number of operational and commercial systems had appeared: Systran, Logos, Meteo, and in particular many Japanese systems. These systems were based in general either on the 'direct' approach to translation, or on the method of syntactic transfer. They relied on bilingual dictionaries sufficient for the text domains in question; linguistic analysis was neither particularly deep or abstract, there was hardly any semantic analysis, and the use of non-linguistic knowledge was entirely absent.

As for research, the dominant framework of MT research until the end of the 1980s was the approach based on essentially linguistic rules on 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 about the domains of texts to be translated. The most notable centre for this research has been Carnegie Mellon University. Nevertheless, these newer knowledge-based systems continued to be essentially rule-based systems, and in any case they remained somewhat of a novelty till almost the end of me decade.

Since 1989 the predominantly rule-based framework 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 1989 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.

This paper will concentrate on these new developments in MT research. It will not describe any one project in detail and projects are mentioned only as examples of trends – there are many others; for further details and for references to the systems mentioned see my recent fuller survey (Hutchins (1)). The paper will also say almost nothing about methods already well established by the end of the 1980s. Furthermore, nothing will be said about the use of commercial systems or the development of aids for translators. The subject is exclusively the development of new methods in MT research. Of course, many of the methods are still experimental and have not yet been tested on a large scale. Nevertheless, the trends are real; since 1989 MT has experienced a reorientation of its methodology sufficient to justify calling the 1990s a genuinely 'new era'.

RULE-BASED SYSTEMS

Before describing these new 'corpus-based' developments in detail I shall begin with rule-based approaches, since here too there have been important theoretical and methodological developments.

Five or six years ago saw the end of two of the most significant transfer-based projects: the Ariane project at Grenoble University and the Eurotra project of the European Communities. These systems exemplified typical features of the so-called 'second-generation' systems: batch processing with post-editing and no interactive components, essentially syntax-oriented and stratificational with three stages of analysis, transfer and synthesis and the processes of analysis and generation passing through series of distinct levels (morphology, syntax and semantics), 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.

Nevertheless, these projects do "live on" to a certain extent in the Eurolang project based at SITE, a French company previously connected with the Ariane project. The project involves collaboration with the German company Siemens-Nixdorf and its Metal system and it is benefiting from experience with Eurotra. The first product of Eurolang is, however, not an MT system as such but a translator's workstation, the Optimizer.

Other transfer-based systems continue in the present decade. There is, for example, the already mentioned commercial system Metal, and the major research at various IBM centres on the LMT ('Logic programming MT') system.

The beginning of this decade saw also the end of some rule-based 'interlingual' research systems: the DLT project in Utrecht based on Esperanto as interlingua, and the Rosetta project at Philips which explored an isomorphic approach to constructing interlingual representations and the integration of Montague semantics. However, major 'interlingual' projects continue to thrive, indeed with even more vigour, particularly in the knowledge-based approach at Carnegie Mellon University. The distinctive features are familiar: a neutral intermediary language for representing the meanings of texts (interlingua) and knowledge databases related to the domain of the texts to be translated. Several models have been developed over the years, and in 1992

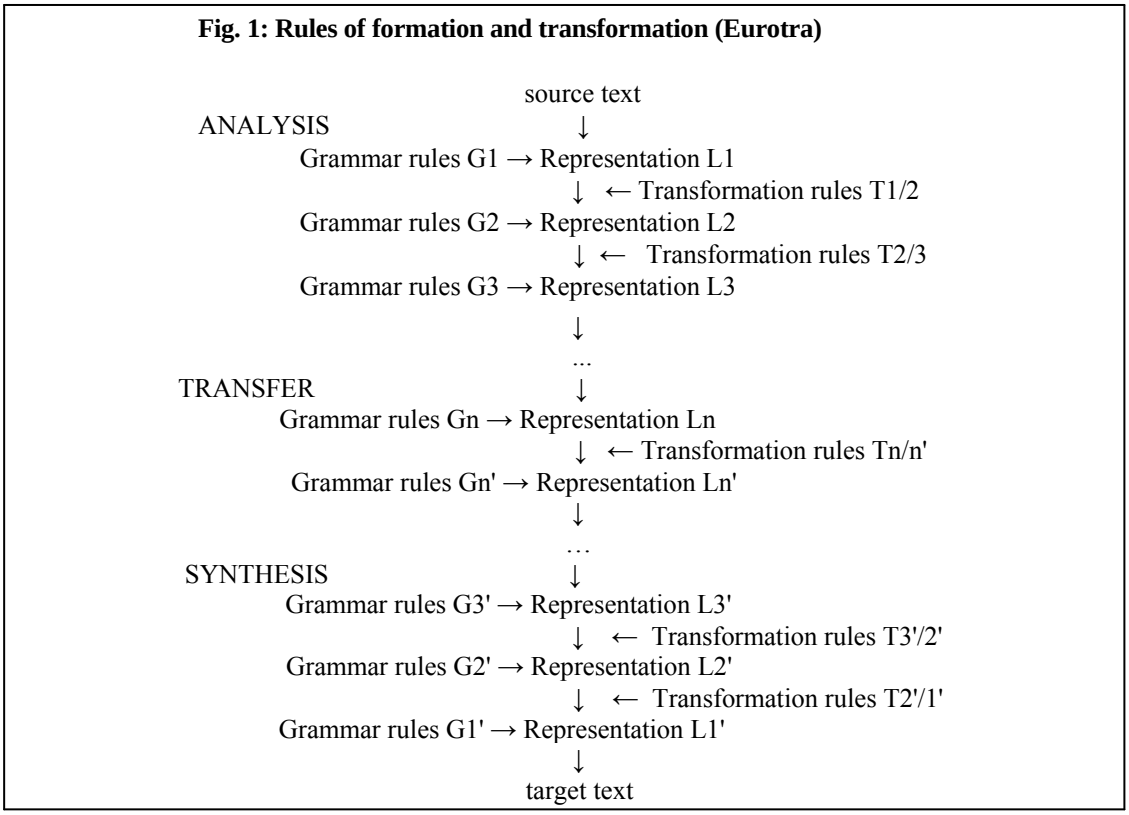
was announced the beginning of a collaborative project with the Caterpillar company with the aim of creating a large-scale high-quality system for technical manuals in the specific domain of heavy earth-moving equipment.

Other 'interlingual' systems are, e.g. the ULTRA system at the New Mexico State University, and the UNITRAN system based on the linguistic theory of Principles and Parameters. There is also the Pangloss project, 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 DARPA. the others being the IBM statistics-based project (see below) 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.

THE 'LEXICALIST' TENDENCY

A characteristic feature of rule-based systems is the transformation or mapping of labelled tree representations. For example (Fig.l), 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.

Fig. 1: Rules of formation and transformation (Eurotra)



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. For example (Fig. 2), to translate English verb *like* into French *plaire* it is necessary to transform the syntactic structure: the English subject *(John)* becomes an indirect object in French, and the direct object *(Mary)* becomes the French subject. These conditions are to be found in the sets of morphological, syntactic and semantic features of the lexical entries of *like* and *plaire.* A slightly more complex set of features is needed to indicate the constraints attached to the English word *likely* and its French equivalent *probable.* The English word requires an infinitival complement, while the French word requires a subordinate clause.

Fig.2: Constraint-based formalism (LFG)

2 (a):

John likes Mary ↔ Marie plait à Jean

like,V:
(↑PRED) = like <SUBJ, OBJ>
(τ↑RED FN) = plaire <SUBJ, OBJ>
(τ ↑AOBJ OBJ) = ↑(SUBJ)
(τ ↑SUBJ) = (↑OBJ)

john, N:
(↑PRED) =john
(τ ↑PRED FN) = jean

mary, N:
(↑PRED) = mary
(τ ↑PRED FN) = marie

F-structure of target language:

$$\begin{bmatrix} \text{PRED} & \text{plaire} \\ \text{SUBJ} & [\text{PRED marie }] \\ \text{AOBJ} & [\text{OBJ} \quad [\text{PRED jean}] \quad] \end{bmatrix}$$

Student is likely to work ↔ Il est probable que l'étudiant travaillera

likely, A:
(↑PRED) = likely <XCOMP> SUBJ
(↑SUBJ) = (↑XCOMP SUBJ)
(τ ↑PRED FN) = probable
(τ ↑COMP) = τ (↑XCOMP)

probable, A:
(↑PRED) = probable <COMP>SUBJ
(↑SUBJ FORM) = il
(↑COMP COMPL) = que

F-structure of target language:

$$\begin{bmatrix} \text{PRED} & \text{probable} \\ \text{SUBJ} & [\text{FORM il }] \\ \text{COMP} & \begin{bmatrix} \text{PRED} & \text{travailler} \\ \text{COMPL} & \text{que} \\ \text{SUBJ} & [...] \end{bmatrix} \end{bmatrix}$$

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 UNITRAN system already mentioned.

An extreme example of the 'lexicalist' approach is the method known as "shake and bake". There are no longer any structural representations, there are only sets of lexical representations (Fig.3).

Fig. 3: 'Shake and Bake' model

monolingual lexical unit (English):

[12] ORTHO like
SEM E1: like (E1), role (E1, experiencer, X1), role (E1, stimulus, Y1)
ARG0 E1
ARG1 X1
ARG2 Y1

monolingual lexical unit (Spanish):

[13] ORTHO gust-
SEM E2: gustar(E2), role (E2, stimulus, X2), role (E2, experiencer, Y2)
ARG0 E2
ARG1 X2
ARG2 Y2

bilingual lexical entry for *like-gustar:*

SPANISH [13] SEM ARG0 E, ARG1 X, ARG2 Y
ENGLISH [12] SEM ARG0 E, ARG1 Y, ARG2 X

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. Today, unification is a central concept for a large number of linguistic theories, and constraint-based grammars and formalisms have attracted many MT researchers: 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.

generation: the same formalism and the same grammars can in theory be applied in both directions.

Several groups have constructed general NLP systems based on unification and constraint-based grammars, which 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; the PLNLP (Programming Language for Natural Language Processing) system provided the foundation for translation systems involving English, Portuguese, Chinese, Korean and Japanese: 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.

The trend towards lexicalist approaches has had 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 with include large amounts of non-linguistic information, such as in the systems developed at Carnegie Mellon or in the UNITRAN system.

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. 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 databanks used by professional translators. 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.

CORPUS-BASED SYSTEMS

While the new approaches, methods and projects described so far can all be regarded as natural progressions from developments having their origins in research of the 1980s or earlier, the emergence of a wide range of what may collectively be called 'corpus-based' approaches and methods represents a new departure in MT research. It is these developments, above all which justify the view that MT has entered a new era.

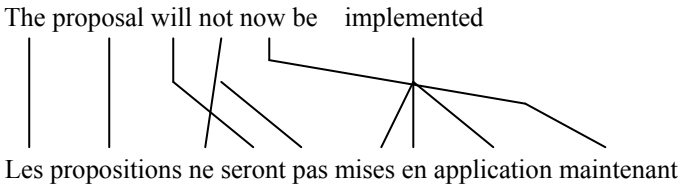
The most dramatic development was the revival of the statistics-based approach to MT in the Candide project at IBM. The major feature is the use of stochastic methods as virtually the sole means of analysis and generation. The IBM research is based on the vast corpus of French and English texts contained 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. The most important point is that this is achieved without using any linguistic information.

It will be seen (Fig. 4) that the method has aligned *proposal* and *proposition, now* and *maintenant.* and *implemented* and the phrase *mises en application.* On the other hand, contrary to linguistic intuition, it has aligned *will* and *seront,* while the word *be* has not been aligned to any French word. On the basis of a large number of such English-French alignments the correspondence and probability frequencies are calculated. The English word *not* corresponds most often to two French words (fertility 2 having a probability of 0.758), and these two words are in general *ne* and *pas* (with probabilities of 0.469 and 0.460); other correspondences are less probable: *non* (0.024), *pas du tout* (0.003), etc. The method was evaluated by translation from English into French.

Fig. 4 Stochastic MT (Candide, IBM)

Alignment:



English: *not*

French	Probability	Fertility	Probability
pas	.469	2	.758
ne	.460	0	.133
non	.024	1	.106
pas du tout	.003		
faux	.003		
plus	.002		
etc.			

What surprised most researchers 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 would like to improve these results, and the IBM group proposes to introduce more sophisticated statistical methods. 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 – the IBM team are true empiricists! Some examples of what is proposed are: (a) the treatment of all morphological variants of a verb as a single word, and (b) the use of syntactic transformations (e.g. *Has the store any eggs? → The store has any eggs QINV; John does not like turnips → John likes do_not_M1 turnips)* to bring the structure closer to that of the target language.

The second major 'corpus-based' approach benefiting likewise from improved rapid access to large databanks of text corpora is what is known as the '**example-based**' (or 'memory-based') approach. Underlying the approach is the basic notion 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 databank 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 example (Fig.5), if a translation is being sought for the English word *fields* a databank might give the following possibilities in French: *domaines, activités, champs.* Each occurrence is given in context. If there is an exact correspondence, e.g. *coalfields → basins-houilliers,* the selection process comes to an immediate end. But if there is no exact match, algorithms must be invoked to find the correct equivalent.

Fig. *5:* Bank of example translations: *field*

English	French
the main fields	les principaux domaines
the following fields	les domaines suivantes
these two fields	ces deux domaines
the specialized fields	les domaines spécialisés
the para-medical fields	activités paramédicales
the magnetic fields	les champs magnétiques
the coal fields	les bassins-houilliers
the corn fields	les champs de blé

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 databanks of actual translations produced by professional translators there is an assurance that the results will be accurate and idiomatic. For example, one of the greatest difficulties of 'rule-based' MT when working from French into English is the selection of the correct equivalent of the preposition *de;* a databank offering a large number of examples could be a major assistance. And there are more complex problems where even greater help could be available, e.g. the translation into French of the phrase *have an effect on* (Fig. 6):

Fig. 6: Example databank for *have an effect on*

English	French
have a direct effect on	ont une influence directe à
have a direct effect on	intéressent directement
have a direct effect on	ont eu une repercussion directe sur
has had a marked effect on	a largement influencé
had a positive effect on	s'est avérée positive dans
had a highly negative effect on X	X en auraient été gravement affectés
will have a decisive effect on	influencera de façon déterminente
would have a detrimental effect on	aurait de fâcheuses répercussions sur

At present, the example-based approach has been used most often to complement more traditional methods based on linguistic rules. However, there are some researchers who contend that the effectiveness of the approach can be fully tested only if it is used as the sole method of generating target text.

A bank of bilingual parallel text can also be used more directly and immediately as a translation tool itself. In this respect, several groups have been developing methods for the alignment of corpora of bilingual texts to provide easily accessible knowledge banks (or 'translation memories') as integral components of workstations for human translators, and

indeed such a feature is already commercially available in the workstations from STAR and TRADOS.

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.** A distinctive feature is the computation of the strengths of links between nodes of networks, and the adjustment of the weightings as a result of actual analyses, i.e. the network 'learns' about links and their strengths for later use. Furthermore, alternatives can be processed in parallel. 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, in an example-based approach by McLean at UMIST. and in the Matsushita transfer-based prototype system.

Connectionism offers the prospect of systems 'learning' from past successes and failures. Previously, learning has meant that systems suggest changes on the basis of statistics about corrections made by users, e.g. during post-editing. This approach is seen in the commercial Tovna system and in the experimental PECOF 'feedback' mechanism in the Japanese MAPTRAN system. A similar mechanism has been incorporated in the NEC PIVOT system .

TEXT GENERATION

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 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 this time, the problem of generating idiomatic output text in the target language was a largely neglected area of MT research. Now, major efforts are now devoted to questions of stylistic improvement of output and to discourse features.

Much of the impetus for this research has come from increasing attention to the need to provide natural language output from searches in databases. 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, not surprisingly, a team based in Montreal long involved in MT. This group has worked on a system for producing marine forecasts in French and English, and on a system for generating bilingual summaries of statistical data on the labour force.

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 not knowing the target language 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 me text to be translated is composed 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.) 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.

CONTROLLED LANGUAGE. DOMAIN-SPECIFIC AND USER-SPECIFIC SYSTEMS

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. Indeed, some of the most successful implementations of MT have been in environments, where the language of input is 'controlled' in some respect. Other systems have been specifically designed for particular subject areas ('sublanguages') or for the needs of specific users. In each case, they 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 sublanguage, 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 are probably the best known examples of controlled language MT. 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 sublanguage is also not new: the well known Meteo has been translating meteorological reports for 15 years. Among the sublanguage systems of recent years there are the CRITTER system for reports on the stock market under development in Montreal, the already mentioned projects at ELU, Pangloss. and the extremely ambitious projects for the development of spoken language translation. The Japanese ATR project has been underway already for seven years and will continue to the end of the century; it is a system for registration at international conferences and for hotel booking by telephone. The European Verbmobil project (Wahlster (2)) is aiming to develop a transportable aid for face to face English-language commercial negotiations by Germans and Japanese who do not know English fluently.

In the past, there were few systems built by users themselves. One example is PAHO (Pan American Health Organization), where two systems were developed for translating from English

into Spanish and from Spanish into English. In the last few years there have been several user-designed systems, typically with restricted vocabularies, for a particular domain and often based on a specific sublanguage. Some of these systems have been developed for 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; Cap Gemini Innovation developed TRADEX to translate military telex messages for the French Army; 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. Such user-designed systems are an encouraging sign that the computational methods of MT and NLP are now spreading more and more 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 could well expand rapidly in coming years.

A NEW ERA

Research on MT has passed through five eras to the present day. The first period began with the memorandum from William 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. As described in this paper, there are now new methods and trends: approaches based on bilingual text corpora, statistical methods, example-based approaches, and new methods using unification and constraint-based grammars. These innovations have all appeared in the last five years and indicate the beginning of a new era for MT. If the direct method characterised the 'first generation' and the indirect methods of transfer and interlingua characterised the 'second generation', what might be the basic features typifying the future 'third generation'?

The general view of many experts is that future systems will combine traditional rule-based methods and the newer statistics-based and example-based methods. They will be hybrid systems. But what kind? In one possible perspective, the 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.

With respect to the base of linguistic rules it may be envisaged that in future hybrid systems:

- rules will be less ambitious and complex than those of indirect systems
- syntactic analysis will be limited to the recognition of surface structures, phrase constituents and dependency relations
- there will be almost no deep analysis of logical relations (quantification, scope of negation)
- semantic analysis will be limited to the identification of roles: agent, instrument, etc.
- lexical information will be extracted mainly from standard sources such as general-purpose dictionaries: consequently the lexicon will include only syntactic categories and perhaps crude semantic features

• fairly simple semantic features will be used for initial disambiguation of input

• rules of lexical and structural transfer will probably apply to shallow representations (although not as crude as in the IBM Candide approach)

• the formalisms will be those of unification and constraint-based grammars.

The corpus-based methods will act to refine and enhance the results and methods, perhaps as follows:

• translation examples stored in aligned bilingual text banks will be used for more delicate disambiguation during source text analysis and for selection of target language equivalents

• statistical information on lexical collocations and monolingual vocabulary frequencies will aid syntactic and semantic analysis of phrases, monolingual disambiguation, and selection of idiomatic target language phrases

• data on probabilities of bilingual equivalences will be used during lexical transfer

• domain-specific knowledge banks will aid monolingual and interlingual disambiguation

• terminological databanks will be used to assist disambiguation of complex phrases and in the selection of target equivalents

• feedback and connectionist methods will be employed to improve grammars (and/or rule bases) and to enhance monolingual and bilingual lexicons

• stylistic features and discourse information will improve output for specific needs and users.

In addition, it can be assumed that many of the newer 'hybrid' systems of the third generation will be directly integrated in general computer-based systems for the production, transmission and management of documents (i.e. more sophisticated workbenches for translators.)

USE OF SYSTEMS

The new research developments described in this paper are taking place against a background of a rapidly expanding marketplace for MT and increasing numbers of users. In recent years, the number of pages translated automatically has increased considerably – at present, more than a million pages annually, or about 300 million words a year (Vasconcellos (2)). The expansion has taken place in large multinational companies and in translation agencies, particularly for the translation of technical manuals. But there has also been an increase in the numbers of non-professional users. Many have purchased cheap PC-based systems, which are certainly crude in linguistic terms. The effectiveness and quality of the systems may be doubtful but the needs of the users are undeniable. To a large extent, MT researchers have not taken up the challenge of designing systems for the non-professional 'occasional' translator. They have also been slow until recently to acknowledge the importance of standards and benchmarks for the evaluation and comparison of the performance, quality and efficiency of commercially available systems.

We can predict an expansion of users of large-scale systems and of users of personal computer systems, and we can also predict an expanding of use of MT systems over electronic networks; in France and Japan, MT is already offered on the PC-VAN, Niftyserve and Minitel networks; in the United States, Systran is available via networks: and CompuServe has just announced an MT service for its users. These are new challenges to MT researchers. What kinds of systems are needed for these new services and demands? We may expect rapid changes in the field of MT in the near future and ultimately the appearance of new systems meeting more closely the actual needs of a wide variety of potential users. Fully automatic

systems capable of producing idiomatic texts comparable to human translation are no longer the goal of MT research. It is now largely focused on the development of systems limited to sublanguages or to specific technical fields.

In favourable conditions, limited-domain systems which are far from perfect can be and are being used successfully and cost-effectively. Of course, everyone wants to see improved quality, but it is not expected in the near future. The new approaches described in this paper have yet to be fully tested in experimental systems, and so it is unlikely that any commercial system based on any methods of the 'third generation' can be expected before the end of the century.

REFERENCE SOURCES

The main sources for information on developments in MT research and for the systems described in this paper are the proceedings of the biennial MT Summit conferences, e.g. in 1991 (3) and 1993 (4), the regular 'theoretical' MT conferences e.g. TMI-90 (5), TMI-92 (6), and TMI-93 (7), and the contents of **MT News International** and **Language Industry Monitor.**

1. Hutchins, W.J. 'Latest developments in machine translation technology.' In: ref (4), pp. 11-34.

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

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

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

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

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

7. **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.