Chapter 9: Strategies and methods since the mid 1960s

9. 1: After ALPAC

In the aftermath of the ALPAC report the reduction of research groups in the United States was dramatic: for a while there were just two groups (at Wayne and Berkeley), until the resumption in 1970 of the research at Texas and the beginning of a new group at the Logos Corporation. In Britain there were now no groups active: the Cambridge team had turned to other interests and the NPL project had ended. Elsewhere the picture was, however, less bleak. In the Soviet Union there had apparently been some reduction of activity (Roberts & Zarechnak 1974), but research continued on both theoretical projects, e.g. the ‘meaning-text’ approach of Mel'chuk (Ch.10.2), and on practical systems, e.g. at the Patent Office in Moscow (Ch.11.5). In France, work was continuing on the CETA project in Grenoble (Ch.10.1), and in Italy, EURATOM began MT investigations (Ch.11.1). In Germany, there appeared a new group at Saarbrücken (Ch.13.2), and Toma continued research which was to lead to the development of the Systran system (Ch.12.1).

In the United States the main activity had been concentrated on English translations of Russian scientific and technical materials. In Canada and Europe the situation was quite different. The Canadian government’s bicultural policy created a “demand for translation which far surpasses the capacity of the market, especially as far as technical translation is concerned” (Chandioux 1977) Shortly after ALPAC the Canadian National Research Council sponsored three English-French projects at Saskatchewan (Ch.12.5), Montreal (Ch.13.1) and Cambridge (Ch.5.2) In Europe the problems of translation were becoming equally acute: in addition to the needs of science, technology, engineering, medicine, and international trade and commerce there were now added the growing demand for translations of administrative, legal and technical documents from and into the languages of the European Communities. As a consequence, the Commission of the European Communities installed Systran in its translation services and inaugurated a multilingual research project (Ch.14).

In contrast to the aspirations of the first decade of research, the goals of MT researchers now became more realistic; no longer were translations expected to be stylistically perfect, the aim was readability and fidelity to the original. At the same time, there now emerged a number of linguistically more advanced systems based on ‘indirect’ approaches to system design and there was an increase in the variety of source and target languages.

9. 2: Direct translation systems

Research has continued on 'direct translation' systems. One was the system developed by the Logos Corporation for the US Air Force which was used to translate American aircraft manuals into Vietnamese (Ch.12.2) Another was the experimental Xonics system, a derivative of the Georgetown system (Ch.12.3), and the recent PAHO system for Spanish and English, also based on the ‘Georgetown’ approach (Ch.12.4). However, the best known was Systran, designed initially as a Russian-English system and later adapted for English-French translation for the Commission of the European Communities (Ch.12.1) Systran may be regarded as essentially a greatly improved version of the Georgetown system; linguistically there is relatively little advance, but computationally the improvements are considerable. The main ones lie in the ‘modularity’ of its programming, allowing for the modification of any part of the processes to be undertaken without risk of impairing overall efficiency, and in the strict separation of linguistic data and computational processes. In this way it avoids the irresolvable complexities of the monolithic Georgetown system (Ch.4.3).

9. 3: Interlingual systems

The most innovative systems have been based on 'indirect' approaches. There were three major projects which investigated 'interlingual' approaches in depth; the Russian-French project by CETA in Grenoble (which had begun some time before ALPAC), the research in Moscow centred on the 'meaning-text' model (which also had its origins before 1966), and the German-English project of the Linguistics Research Center (LRC) at the University of Texas (Ch.10). All were heavily indebted to current linguistic theories. Both the CETA and LRC groups were influenced by the Chomskyan theory of transformational grammar (Ch.3.5). It was argued that while languages differ greatly in 'surface' syntactic structures they share common 'deep structure' representations, which may be regarded as forms of 'universal' semantic representations. SL analysis passed through various levels ('strata') from 'surface' structures to 'deep' structures; and TL synthesis from 'deep' structures to 'surface' representations. In this conception, both groups were influenced also by Lamb's stratificational theory. At both CETA and LRC, interlingual representations were abstract formulas in terms of 'logical' predicates and arguments. However, in neither case were lexical items converted into interlingual representations, there was no attempt to decompose lexical items into semantic primitives. In brief, the CETA and LRC systems were interlingual in syntax, but not interlingual in semantics. By contrast, the 'meaning-text' model developed by Mel'chuk and others (Ch.10.2) was an attempt to design a fully interlingual system. As in CETA and LRC, there were various levels of analysis and synthesis, from surface forms to 'deep' syntactic forms; but in addition there was a 'deep' semantic level where all lexical synonymy and ambiguity was resolved.

While CETA and LRC were quite thoroughly tested, the Soviet system was not implemented, mainly for external reasons. Even if it had been, it is doubtful whether it would have been more successful than CETA and LRC, neither of which went beyond the experimental stages. The basic fault was the rigidity of the processes: a failure at one stage of analysis corrupted all subsequent stages, and too often, too many structures were produced for a sentence which then had to be 'filtered out' later. It was largely as a consequence of disappointing results that the Grenoble group adopted a basically 'transfer' design for its new GETA system (Ch.13.3), and that when the Texas group resumed MT research in 1978 it also adopted an essentially transfer strategy (Ch.13.4)

9. 4: Transfer systems

In retrospect, these 'interlingual' systems were perhaps too ambitious at the time; the more cautious 'transfer' approach was probably more realistic as well as being more flexible and adaptable in meeting the needs of different levels and 'depths' of syntactic and semantic analysis.

In broad terms, the 'transfer' systems may be divided into those based on the MIT-type 'syntactic transfer' pattern and those in which syntactic analysis generally went further than 'surface' structures and more semantic analysis was incorporated. Examples of the former were the experimental Russian-English project at EURATOM (Ch.11.1) and the Moscow Patent Office system which was installed for translating American patent abstracts into Russian (Ch.11.5) The examples of the latter type were all to emerge since the mid-1960s. They included the POLA system at Berkeley

(Ch.11.2), Kulagina's system at MIAN (Ch.11.5), and the TAUM project at Montreal (Ch.13.1). The latter system may been seen in retrospect to have inaugurated the development since the early 1970s of techniques which have resulted in the advanced GETA system at Grenoble (Ch.13.3), the SUSY system at Saarbrücken (Ch.13.2), and the Eurotra project of the European Communities (Ch.14.2).

In these newer transfer systems, the goal of analysis was the production of SL representations which resolve the syntactic and lexical ambiguities of the sentence in question, without necessarily providing unique representations for synonymous constructions and expressions (Ch.3.9) Thus, while it might resolve homonyms such as *watch*, and equate lexical synonyms such as *expensive* and *costly*, it would probably not resolve certain prepositional phrase ambiguities or equate *buy* and *sell* constructions (Ch.3.6) In comparison with 'interlingual' approaches there was more conservation of information about SL sentence forms and also greater reliance on bilingual dictionaries for providing information on SL-TL structural changes during the transfer stages. In sum, the 'depth' of analysis in transfer representations was less than in the representations of 'interlingual' systems.

One of the major theoretical justification for 'transfer' systems is their extensibility to further language pairs in a multilingual environment (Ch.3.9) Nevertheless, there are sufficient advantages in the inherent 'modularity' of such designs to warrant their adoption even when only a one-directional bilingual system is envisaged. The TAUM system was developed specifically for English-French translation, the POLA for Chinese-English, and the MIAN system for French-Russian. The SUSY and GETA systems were developed initially for a particular pair (Russian-German and Russian-French respectively), but have now been extended to other languages; and Eurotra has been designed *ab initio* as a multilingual system.

9. 5: Semantics-based systems

At the same time as the more advanced 'transfer' systems were beginning to appear, MT research saw the introduction of the first semantics-based approach, inspired by methods and techniques of Artificial Intelligence (AI). This was the experiment by Wilks at Stanford University on an 'interlingual' English-French system.

Since the mid-1960s MT systems have been predominantly syntax-based. This applies as much to the 'interlingual' and 'transfer' approaches as it does to the earlier systems. However much semantic information is included in interlingual and transfer representations, syntactic analysis is the central component: semantic features are attached to syntactic structures and semantic procedures operate primarily after syntactic structures have been determined. These systems are syntax-based in another sense: their analysis and synthesis procedures are restricted almost exclusively to sentences. Features of discourse cohesion across sentences, such as pronominalisation and topicalisation (Ch.3.7), have been neglected. The deficiencies have been recognised by many researchers, e.g. at LRC, TAUM and GETA, but there are still only tentative suggestions for handling intersentential relations.

The importance of Wilks' MT research and of other AI projects which have followed, such as the Yale projects (Ch.15.2), has been precisely in the exploration of semantics-based approaches to language analysis. Basic components of AI approaches are: semantic parsing, i.e. analysis of semantic features ('human', 'concrete', etc.) rather than grammatical categories; lexical decomposition into semantic networks (e.g. Schank's conceptual dependency model); and resolution of ambiguities and uncertainties by reference to knowledge bases. Initially, AI approaches were pursued as alternatives to ‘traditional’ syntax-based systems, but recently MT researchers are experimenting with AI methods in combination with linguistic analysis techniques, e.g. TRANSLATOR (Ch.15.7) and some Japanese systems (Ch.18.12)

9. 6: Interactive and restricted language systems

One consequence of the ALPAC report was to convince many (if they were not already by Bar-Hillel) that fully automatic translation of good quality is unattainable. Some have explored the possibilities of interactive MT, in which man and machine collaborate in the production of translations; some have argued for systems based on restricted vocabulary and syntax; and others have exploited the computer’s capacity to retrieve information from large databases by developing mechanized dictionaries to aid translators.

Interactive systems appeared in the late 1960s, with the experimental MIND project at RAND and the Chinese-English CULT system in Hong Kong (Ch.17.8-9). As on-line processing became more generally available during the late 1970's, commercial organizations started to investigate the possibilities and as a result, in recent years systems have appeared on the market place (Ch.17) The amount of human interaction varies considerably from one system to another. Nearly all require human operators to assist in the analysis of SL text and in the compilation of dictionaries.

Imposing restrictions on the vocabulary and syntax of input texts remains attractive (Ch.17.1-4). It is sometimes combined with interactive facilities, as in TITUS and Smart, and sometimes implemented on fully automatic systems, as in the case of Xerox’s use of Systran.

9. 7: The developments after the mid 1970s

In the years since ALPAC the development of automatic dictionaries and terminology databanks has contributed in large part to the now increasing acceptability of MT. Until the 1970s practising translators were largely excluded from involvement in computer-based translation. MT was predominantly an affair for academics and government scientists. This has been changed by the development of machine aids for translators (Ch.17.6). They have been created in response to the ever urgent needs of translators, particularly in large governmental and industrial units, for rapid access to up-to-date glossaries and dictionaries of terminology in science, technology, economics and the social sciences in general. The potential of mechanized dictionaries was recognized from the very beginning of MT research, e.g. by Booth and by Oettinger, but it has only been realized as a result of recent technological advances, particularly in the availability of on-line access via public telecommunication networks and in the development of text editing and word processing equipment.

Equally important for the changing attitude to MT has been the decision by the Commission of the European Communities in the mid-1970s firstly to test and develop the Systran system for use in their translation services (Ch.14.1) and then subsequently to approve a multinational MT research project for the development of the multilingual Eurotra system (Ch.14.2) At the same time, the Commission organized a conference in 1977 under the title *Overcoming the language barrier* at which most contemporary MT groups presents accounts of their research (CEC 1977) In subsequent years the Commission has supported Aslib, the British organization for special libraries and information services, in the organization of a series of annual conferences which since 1978 have successfully drawn the attention of translators and business organizations to developments in MT and related activities (Snell 1979; Lawson 1982; Lawson 1985)

In retrospect the mid-1970s may prove to be a turning point for MT. In the Soviet Union there was a similar renewal of interest. In 1975 for the first time since the mid-1960s an international MT conference was held in Moscow. It was organized by the All-Union Centre for Translations which since 1974 has assumed responsibility for the development of MT systems in the Soviet Union (ch.18.3-6). The conference was the first of a series held every four years. In the United States, the MT revival has been much more fitful. In March 1976 a seminar was organized in Rosslyn, Virginia, for the U.S. government Foreign Broadcasting Information Service (FBIS 1976). The FBIS was translating annually over a million words and was expecting demand to grow; it was "facing translation problems" and wanted to "reassess the state of the art" of MT and machine aids for translators. However, the general mood was pessimistic. Petrick, in his summary, concluded that "currently operational or projected machine-translation systems are only marginally different in their underlying organization and design than their predecessors" and "I would not expect any current MT systems to compete with human translation except where low quality, unedited output suffices." On the other hand, Hays saw a future for interactive MT systems with text editing facilities.

In the event, US involvement has continued to be low. For many years only the University of Texas has been involved in fully automated systems (and they have been funded by a non-US company); however, in recent years there has been some increase in research interest elsewhere in the US, indicated by the conferences in August 1985 at Georgetown University (*LM* 25, Oct 1985) and at Colgate University (Nirenburg 1985). Much of the renewed 'academic' interest is centred on AI approaches (Ch.15), and is small-scale experimental research. Since the mid 1970s the main US activity has been in the commercial developments of interactive MT systems, where there has been considerable progress, although these systems are often marketed as adjuncts of office automation systems, and avoid the tainted name 'machine translation' in favour of 'computer assisted translation.'

9. 8: Computational and linguistic techniques and issues

Whereas in the early 1960s, the main advances in language processing took place within the context of MT research, since 1966 most innovations have occurred outside this environment. In part this was a result of ALPAC's recommendation that support should concentrate on computational linguistics rather than MT projects. However, there was already a considerable growth of natural language processing activity outside MT before ALPAC, notably on question-answering and information retrieval systems. Above all, artificial intelligence research was turning increasingly towards problems of language understanding and text analysis, e.g. Weizenbaum's ELIZA, Winograd's SHRDLU, Schank's conceptual dependency model, etc. (Boden 1977)

The particular relevance of AI methods and techniques will be discussed later (Ch.15 and 19.3) This section summarizes relevant issues and developments of the late 1960s and early 1970s in general parsing and analysis techniques.

9. 9: Separation of linguistic data and algorithms

In the earliest systems, programs included procedures which combined linguistic information and specific actions. For example, the ITMVT English-Russian program (Ch.6.1) consisted of a succession of searches for specific English words; and the Georgetown algorithm (Ch.4.3) was driven by specific lexical items. The advantages of separating algorithmic procedures from linguistic data (and not only lexical but grammatical as well) was argued particularly by Yngve of MIT (Ch.4.7) and Lamb of Berkeley (Ch.4.10). The result was a tripartite design for parsing: a dictionary (with grammar codes), a processing algorithm, and a store (table) of grammar rules. It enabled the linguists to concentrate on linguistic analysis, and programmers to concentrate on efficient algorithms; and it meant that rules could be changed or modified without changing the processing algorithm. Garvin (1972), on the other hand, argued that computationally more efficient programs were possible in a 'bipartite parser' where the algorithm searches for linguistic patterns as directed by grammar rules embodied in the algorithm itself. Although there is general agreement on separating lexical data from algorithmic routines, there are still differences of opinion regarding the desirability of separate grammar rule tables and abstract algorithm routines (see, e.g. Ch.13.4 on the latest Texas system.) In the 1970s, however, this approach was widely adopted. It is seen most clearly in the development of 'abstract' tree transduction procedures (below)

9. 10: Modularity

Modular design is most apparent in the subdivision of syntactic analysis programs. During the 1960s a number of projects developed parsers which consisted of a series of 'passes' to identify groups and phrase structures (noun phrases, verb groups, clauses, etc.); a typical example was the 'fulcrum' analyzer of Garvin at Ramo-Wooldridge (Ch.4.6) The disadvantage of the monolithic complexity of the Georgetown system was that particular procedures (e.g. to deal with adjective-noun congruence) could not be changed without affecting other parts of the system. The advantage of dividing procedures into relatively independent subroutines is that parts can be modified and tested without running such risks (and, of course, such modularity is easier if grammar rules and algorithms are separated). Since the mid-1960s modularity has become the norm, even in systems of the 'direct translation' type such as Systran (Ch.12.1).

9. 11: Transformational parsing

Given Chomsky's demonstration of the inadequacies of phrase structure grammars (Ch.3.5) and the disappointing experience (both within and outside MT) with parsers based on equivalent formalisms (e.g. dependency parsers and predictive analyzers), it was natural that researchers should investigate transformational parsers. There was a further reason why the transformational theory was attractive. This was the claim that while languages may differ considerably in 'surface' structures they all share the same 'deep structures' (Chomsky 1965:117ff.). The theory seemed to offer a way of dealing with syntactic equivalencies between languages, and it proved a stimulus to research on 'interlingual' approaches to MT, as already indicated.

It was soon found, however, that the Chomskyan formulation of transformational rules could not be easily implemented in a syntactic analysis program. In fairness to Chomsky, it needs to be pointed out that he has never suggested that his model could be applied in computational systems for language analysis; indeed, he has written in relation to the intellectual climate of the 1950s: "As for machine translation and related enterprises, they seemed to me pointless as well as probably quite hopeless" (Chomsky 1975: 40).

The basic reason why transformational analysis does not work easily is that Chomskys model is conceived as a generative grammar; it accounts for structures by describing how they may be formally derived from an initial node S by rules such as S → NP + VP, NP → A + N, VP → V + NP, etc. and by transformational rules which convert and merge phrase structures. It does not describe how structures can be recognized; i.e. it does not supply a mechanism for parsing. Researchers such as Petrick (1971) and the MITRE group (Grishman 1976, King 1983a) discovered that parsers based on procedures with reverse transformational rules are inordinately complex; many alternative sequences of transformational rules may have applied in the generation of any surface structure; each possibility must be tried and each potential 'deep structure' must be tested for well-formedness. Furthermore, many transformational rules, such as those forming coordinate structures (Ch.3, fig.5), delete deep structure information, and there is no way this information can be reconstructed with certainty. Much ingenuity was devoted to transformational parsing, but the general conclusion is that the methods work well only on very restricted domains (King 1983a).

9. 12: Filtering

Many analysis programs incorporate 'filters' which check for well-formedness of derived structures. For example, a surface structure analyzer determines which sequences of grammatical categories (assigned by dictionary lookup or morphological analysis) are legitimate parsings. A 'deep' structure parser then tests each putative surface structure, either in terms of possible 'deep structure' sources (as in transformational parsing) or in terms of semantic coherence. The semantic criteria may include 'selection restrictions' or valid 'case' relationships (Ch.3.6 and below) Such a model of analysis adopts, therefore, a multi-level or 'stratificational' conception of language structure: morphological, syntactic, semantic (Ch.3.10); and each stage of the analysis program acts as a 'filter' for structures accepted by the previous stage. The basic difficulties of the 'filtering' approach are the risks, on the one hand, of rigidity and, of eliminating unusual (but valid) structures and, on the other, of allowing too many structures of doubtful validity. Both faults were present in the analysis programs of CETA and LRC (Ch.10).

9. 13: Augmented transition network (ATN) parsing

The ATN parser developed by William Woods (1970, 1973) was the outcome of research he undertook with Kuno in the Harvard Computation Laboratory on improvements to the 'predictive syntactic analyzer' (Ch.4.9) Woods' parser differs in two important respects from the finite state 'grammar' of the Harvard analyzer. Firstly, the arcs of one finite state network may be labelled with the names of other networks; thus, in the extremely simple 'grammar' of three networks displayed in Fig. 12 below, transition to state 2 requires the first word of a sentence (S) to be an aux(iliary verb), while transition to state 1 or from state 2 to 3 requires the satisfactory completion of the NP network, i.e. testing for the categories 'pron(oun)', 'det(erminer)', 'n(oun)' and reaching state 7 or state 8. The optional PP network – its optionality indicated by an arc looping back to the same state – requires the testing for a 'prep(osition)' and again the satisfactory completion of the NP network. As such, this parser would still be no more powerful than a phrase structure grammar. It can in fact be made equivalent to a transformational grammar.

Its 'transformational' capability is achieved by adding tests and conditions to the arcs and by specifying 'building instructions' to be executed if the arc is followed. Thus, for example, transition of arc 'aux' to state 2 would specify the building of the first elements of an interrogative (phrase) structure, which could be confirmed or rejected by the conditions or instructions associated with other arcs. Likewise, the transition of an arc recognizing a passive verb form would specify the building of elements of a passive construction to be confirmed or rejected as later information is acquired. As a consequence, Woods' parser overcomes many of the difficulties encountered by previous researchers in attempting to devise parsers with reverse transformational rules (9.11 above).

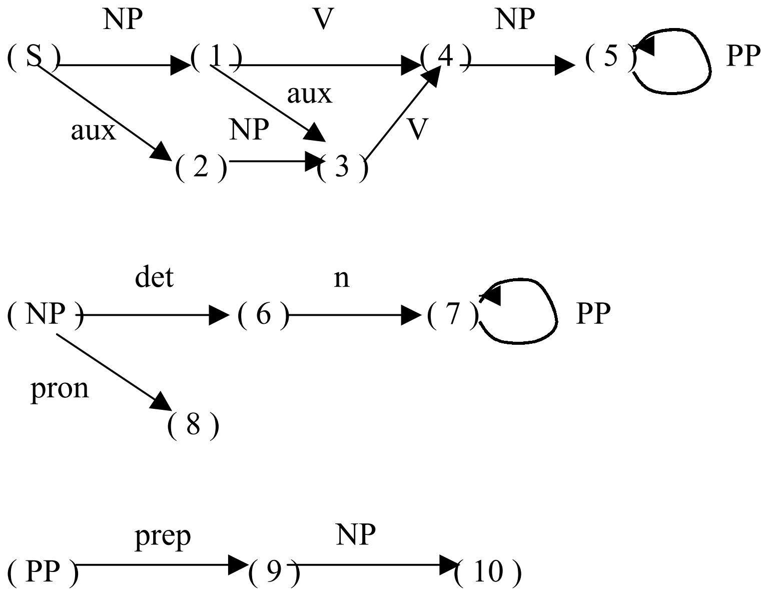


Fig.12: Partial ATN grammar

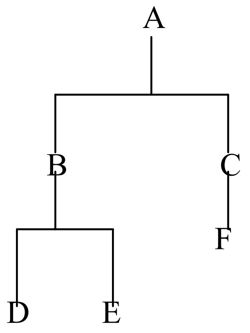
One of the principal attractions of ATN parsers is that they are by no means restricted to syntactic analysis. Indeed in AI systems they are commonly used for deriving semantic representations (e.g. Simmons 1973). Conditions may specify any type of linguistic data: thus, arcs can test for morphological elements (suffixes and verb endings) and for semantic categories ('animate', 'concrete', etc.); and instructions can build morphological analyses and semantic representations. Furthermore, because the arcs can be ordered, an ATN parser can make use of statistical data about the language and its grammatical and lexical structures.

Normally ATN parsers operate top-down (cf. Ch.3.4), with all the disadvantages that entails, principally in wasteful reiterated analyses of lower level constituents. However, it is also possible for ATN parsers to be implemented breadth-first, exploring all possible paths 'in parallel', and thus minimising backtracking routines (see Johnson 1983 for an introduction to ATN parsing.)

9. 14: Tree transducers

Analysis and synthesis programs involve the transformation of one structure (e.g. string of elements or phrase structure tree) into another. In earlier systems (and in many later 'direct' systems) such transformations were implemented by routines specifying both constituent elements and their interrelationships. As a consequence of separating linguistic data from algorithmic data, there evolved the notion of 'abstract' tree transduction algorithms. Within MT research, they appeared first in the Q-system of the TAUM project (Ch.13.1).

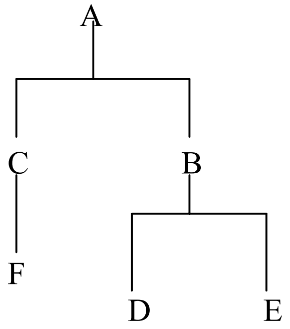
Such algorithms are based on the fact that any tree can be expressed as a string of bracketed elements, thus the tree:



can be expressed as: A(B(D,E),C(F))
The conversion of one tree into another is a matter of defining rewriting rules applying to the whole or part of a string (tree), e.g.

A(B(*),C(*)) → A(C(*),B(*))

where * indicates any subtree or subordinated element. This would convert the tree above into:



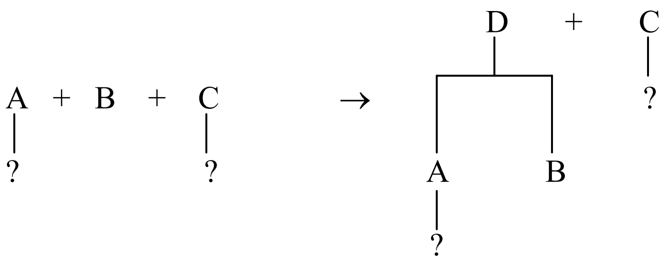
Tree-transducers are able to deal with the occurrence of optional elements in trees or sub-trees which are not affected by the conversion rules, e.g. the occurrence of an unspecified string or tree '?' between B and C at the same level. (For example, B and C might be elements of a phrasal verb *look...up*.) The rule might then be written:

A(B(*),?,C(*)) → A(C(*),?,B(*))

They can also be applied in the conversion of strings of elements or subtrees into trees or into other strings of elements or subtrees, e.g.

A(*)+B+C(*) → D(A(*),B)+C(*)

i.e.

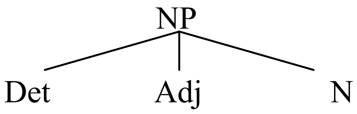


Tree-transducers can therefore be applied not only in syntactic analysis and in the transfer components of MT systems, which are clearly their most obvious applications, but also in procedures involving strings and loosely structured representations (e.g. in morphological analysis, cf. example in TAUM, Ch.13.1). Applied in syntactic analysis tree-transducers operate as bottom-up parsers (Ch.3.4), building upwards from strings of grammatical categories to phrase structure trees (NP or VP) and to sentence structure trees.

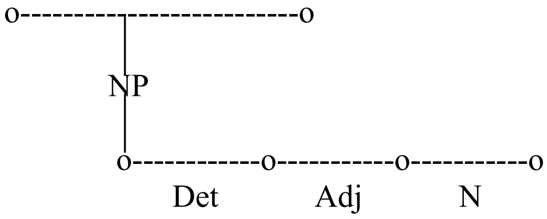
9. 15: Charts

Natural language analysis involves normally the computation of a large number of possible partial analysis, many of which, notoriously, lead to dead ends. (This was amply demonstrated in the finite state and phrase structure parsers of the first MT decade.) Charts have been developed as a means of keeping track of partial analyses and provide the data for the selection of the 'best' parsings (Varile 1983).

The basic principal of the chart is that tree structures can be represented as labelled graphs. For example, the tree:



(where the dominance relations are explicit and the linear (precedence) relations are implicit, cf. Ch.3.4). As a chart this appears as:



(where the linear relations are explicit and the dominance relations implicit)

Alternative analyses can be easily accommodated. Suppose, for example, two analyses of *The peasants were revolting*: one in which the VP (*were revolting*) is interpreted as Cop + Adj, and the other in which it is interpreted as Aux + PrPt (i.e. a present durative form). The two interpretations can be included on the same chart (Fig.13):

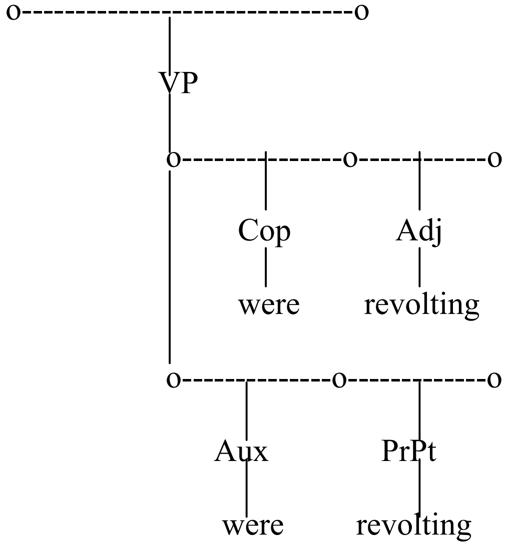


Fig.13: Partial chart representation I

The interpretation of charts in parsing involves the creation of arcs which subsume sequences of arcs, i.e. the replacement of complete analyses for partial analyses. In the process, some arcs for partial analyses are not subsumed, they represent failed analyses. For example, the sentence *They were condemned prisoners* might start from the chart in Fig.14:

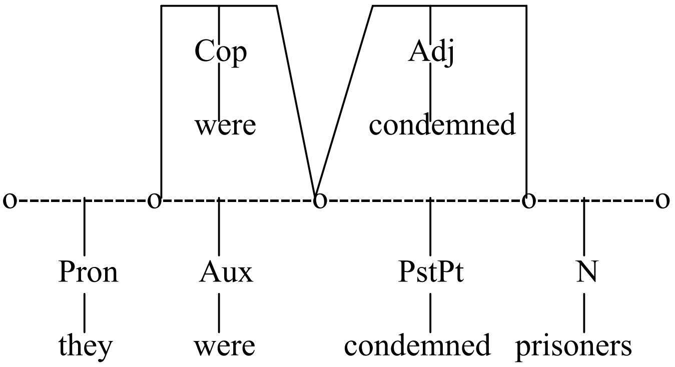


Fig.14: Partial chart representation II

The next stages are illustrated in Fig.15: an arc enclosing Adj(condemned) and N(prisoners) is drawn, i.e. NP (Adj(condemned), N(prisoners)); and an arc enclosing Aux(were) and PstPt(condemned) is also be drawn, i.e. the passive construction: VP(Aux(were), PstPt(condemned)):

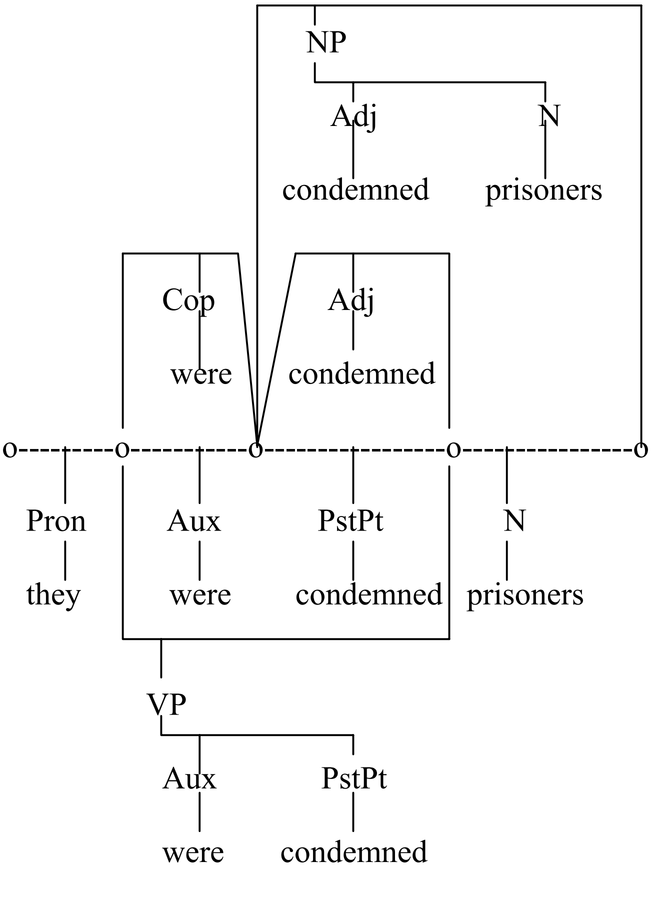


Fig.15: Chart representation III

The parser now looks for complete stuctures. In the lower part of the chart, there is no arc which can subsume VP and N; there is one which could subsume Pron and VP but this would leave N(prisoners) unaccounted for. As a result the only successful parse joins Cop and NP as a VP, and then Pron and VP as S, producing the final complete arc: S (NP (Pron(they), VP(Cop(were), NP(Adj(condemned), N(prisoners)))

Charts made their first appearance in MT systems in the TAUM project (Ch.13.1), and in the context of the MIND project (Ch.17.8), where they were developed by Martin Kay and Ronald Kaplan. Subsequently, charts have been used in GETA, SUSY, METAL and Eurotra. The flexibility of the chart method of handling parsing data is indicated by the fact that they have been used with a number of different types of parsers.

9. 16: Semantic analysis

Before the late 1960s the analysis of sense relations among lexical items in sentences and texts tended to be divorced from the analysis of syntactic relations. The thesaurus approach of the Cambridge group and the correlational analysis of the Milan group were conceived essentially as alternatives to syntax-based approaches. Semantic information was incorporated in word-for-word and syntax-oriented systems in the form of semantic features (e.g. 'human', 'mass', 'quality') attached to lexical entries in order to resolve ambiguities of structure and to aid selection of TL forms. The use of semantic features and 'selection restrictions' has remained a standard method since the 1960s (Ch.3.6).

Analysis of semantic relations as such was not practised until the appearance of 'interlingual' systems. These systems (e.g. CETA, LRC) introduced the analysis of logical relations (predicates, arguments, attributes), which has continued to be employed in later 'transfer' systems and in other MT approaches. On the other hand, the analysis of lexical sense relations (synonymy, paraphrase, causation, etc.) and lexical decomposition has generally appeared only in systems adopting AI approaches (Ch.15 and below)

The most popular type of semantic analysis has been the use of 'case frame' analysis, adopted in many systems ('direct' as well as 'transfer') during the 1970s and now established as a standard proven technique. The notions of 'case' and 'case frame' were introduced during the late 1960s primarily from the work of Charles Fillmore (1968). The intention was to capture the equivalence of semantic roles in sentences such as:

John sold the car to Mary
The car was sold to Mary by John
Mary was sold the car by John

where it can be said that in each one *John* is the 'agent' of the transaction, *Mary* the 'beneficiary' (or recipient) and *the car* the 'object' (or 'patient'). Agent, Beneficiary and Object (or Patient) are case relations; others are Instrument, Source, Goal, etc. In English they are often expressed by prepositions (*by* for Agents, *to* for Beneficiaries, *with* for Instruments, etc.); in other languages by 'surface' case markers (e.g. in Latin or Russian: accusative, dative or ablative case endings). The 'case frame' for a verb is the specification of those case relations which may (optionally or obligatorily) occur with it. In this respect case frames are an extension of the notion of valency (Ch.3.4). Cases are widely assumed to be 'universal' and language-independent; but there is remarkably little agreement on what the 'universal cases' are! Although originally formulated in the context of generative grammar, cases and case frames have been applied in many different contexts; they are particularly popular in Japanese MT projects (Ch.18.2) and in AI approaches, where cases are frequently basic constituents of semantic grammars.

9. 17: Semantic parsing.

Parsers and analysis programs in MT, as in most computational linguistic processing, have been based on the sentence as the unit of analysis. Syntactic theory has developed primarily as a description of sentence structure (cf. Chomsky's definition in 3.5 above). It does not account for the choice of specific sentence structures in context, e.g. selection of a passive instead of an active, nor for relationships between sentences within texts, e.g. selection of pronouns to refer to antecedent nouns. And it is not only syntactic analysis that has been confined to sentence structure. Most of semantic analysis has been restricted to resolving ambiguities within sentences.

Semantic parsing was developed within the context of AI research as a means of overcoming the well known inadequacies of sentence-based syntactic parsing and as a method of deriving representations in language understanding systems. In its ‘pure’ form, semantic parsing is based on the recognition of semantic features (of lexical items in text) either in sequences and patterns or in ‘conceptual’ frameworks. An example of the former is Wilks’ parser which looks for semantic ‘templates’, such as the sequence of ‘primitives’ MAN HAVE THING (Ch.15.1) An example of the latter is Roger Schank’s conceptual dependency approach in which the occurrence of a lexical item activates predictions of semantically (or conceptually) compatible items. Thus ‘drink’, as an item of INGEST(ation), predicts an actor, a liquid object and an optional container, within a specific network of relationships (cf.Ch.15.2). Both approaches may be regarded as extensions of the ‘case frame’ notion beyond the sentence, since both templates and predictions can act across sentence boundaries. Although semantic parsing has been employed most frequently in AI approaches to MT, the basic notions are to be seen incorporated increasingly in otherwise essentially syntax-based systems (cf. GETA and Eurotra).

9. 18: Outline of project descriptions and general sources

The next nine chapters deal with the many MT systems which have been developed since the mid-1960s. The arrangement is primarily according to system design, with chronological divisions also applied as appropriate. Interlingual systems immediately after ALPAC are described first (Ch.10), followed by other ‘indirect’ systems of the period (mainly of the ‘syntactic transfer’ type) in Ch. 11. The ‘direct translation’ systems since the mid-1960s are treated in the next chapter: Systran, LOGOS, PAHO and others. This chapter is followed by extensive treatments of the important ‘transfer’ systems TAUM, SUSY, GETA and METAL, and also the more recent Logos system and Czech research. Chapter 14 deals with the systems associated with the Commission of the European Communities, the various Systran systems and the Eurotra project. The increasingly significant work on Artificial Intelligence approaches is described in the next chapter. This is followed by accounts of various other ‘interlingual’ projects since the mid-1970s; and in Ch. 17 by descriptions of restricted language and interactive MT systems. Lastly, Ch. 18 describes research in the Soviet Union since 1975 and the recent important MT activity in Japan. The final chapter of the book attempts to summarize the present situation and point to possible future prospects.

The main bibliographic sources for the systems are given in the appropriate sections. Major sources for the period as a whole are the surveys by Bruderer (1978), Hutchins (1978), Josselson (1971), Kulagina (1976), Lehmann & Stachowitz (1972), Locke (1975), Roberts & Zarechnak (1974), Slocum (1984a), Tucker & Nirenburg (1984), and Whitelock & Kilby (1983). All contain substantial bibliographies (particularly Bruderer's handbook), but in addition the excellent bibliography by Mel’chuk & Ravich (1978) covering the period 1964-70 must be mentioned.