[From: Internat.conf. *Methodology & Techniques of Machine Translation*, British Computer Society, Cranfield, 13-15 February 1984]

METHODS OF LINGUISTIC ANALYSIS IN MACHINE TRANSLATION

J.Hutchins
(University of East Anglia)

Abstract *This paper is concerned with linguistic aspects of machine translation (MT), and specifically with the general methods that have been adopted for the analysis of texts. Most emphasis is given to techniques of syntactic analysis and to problems of semantic analysis. No full description is given of any particular MT system; but illustrations are taken from some of the most important past and present systems and projects.*

1) General strategies.

In broad terms, there have been three types of overall strategy adopted in MT systems for the translation of texts from one language into another language (Hutchins 1978).

The first approach, employed in nearly all MT systems until the late 1960's, was the 'direct translation' approach (fig.l). Systems were designed in all details specifically for one particular pair of languages, in most cases for Russian as source language (SL) and for English as target language (TL). The basic assumption was that the vocabulary and syntax of SL texts need not be analysed any more than strictly necessary for the resolution of ambiguities, the correct identification of appropriate TL expressions and the specification of TL word order. Thus if the sequence of SL words produced an acceptable sequence of TL words when converted one at a time, then there was no need to identify the syntactic structure of the SL text. In its crudest form this approach is seen in the early word-for-word systems in which essentially each word of the SL text was substituted by a word or selection of words in the TL; occasionally some rearrangement of TL words was attempted, but the onus of providing a readable text rested on a post-editor who had to select the correct translation from the choices offered and to put the text into grammatical TL sequence. Later examples of the 'direct' approach incorporated some syntactic analysis and attempted to select the correct TL equivalents of SL input words. Most, however, still required a good deal of post-editing for acceptable results.

By the early 1960's it had been recognised that MT could not advance much further without greater sophistication in linguistic analysis, particularly with respect to the fundamental semantic problems of translation. The achievements of theoretical linguistics at this time seemed to promise considerable improvements in MT systems. Research was begun on 'indirect translation' systems in which SL text analysis and TL text generation are kept separate and conversion is achieved via an 'interlingual' representation or a 'transfer' component operating on abstract intermediary SL and TL representations.

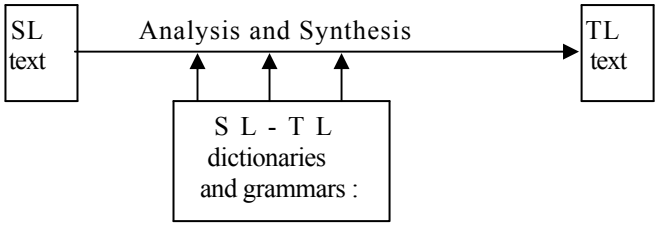


Fig.1. 'Direct translation' system

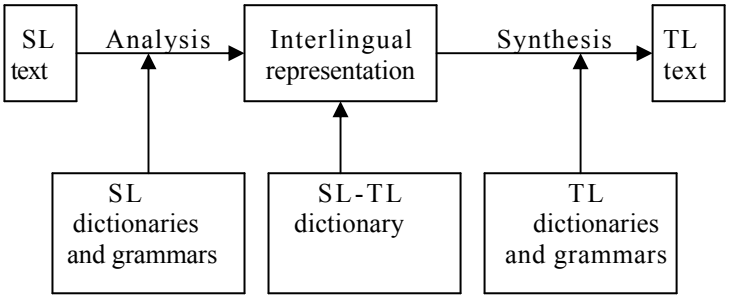


Fig.2. 'Interlingual' system

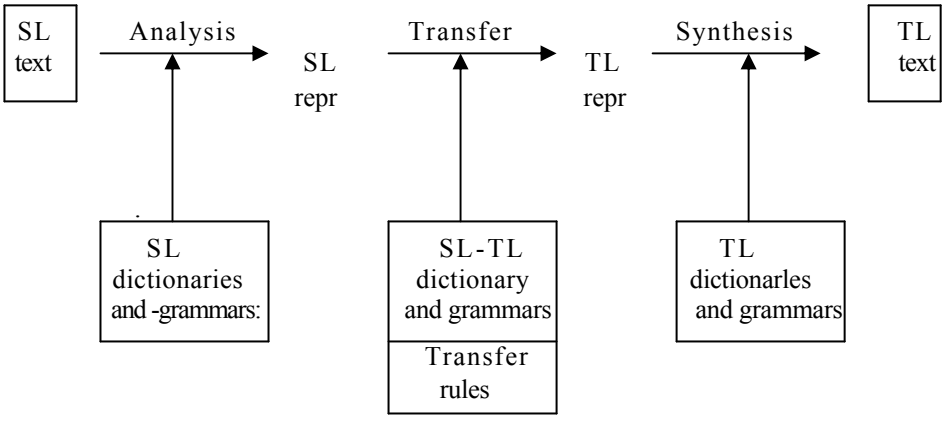


Fig.3. 'Transfer' system

The 'interlingual' approach assumed the possibility of converting SL texts into semantico-syntactic representations common to more than one language. From such 'interlingual' representations texts would be generated into other languages (fig.2). In such systems translation from SL to TL is in two stages: in the first stage SL texts are fully analysed into interlingual representations, and in the second stage interlingual forms are the sources for producing (synthesising) TL texts.

The third approach to overall MT strategy is the 'transfer' approach, which is favoured in most current research (fig.3). Rather than operating in two stages through a single highly abstract interlingual representation, there are three stages involving underlying representations for both SL and TL texts; i.e. the first stage converts SL texts into SL 'deep' representations, the second converts these into TL 'deep' representations, and the third produces from these the final TL text forms. Whereas the interlingual approach necessarily requires complete resolution of all ambiguities and anomalies of SL texts so that translation should be possible into any other language, in the 'transfer' approach only those ambiguities inherent in the language in question are tackled. For example the English verb *know* may be translated as either *connaître* or *savoir* in French, but this does not mean that *know* is ambiguous in English; there is no need when analysing English text to determine which kind of 'knowing' is involved. The 'interlingual' approach would require such analysis, the 'transfer' approach does not; problems of mismatch between SL and TL 'deep' representations are resolved in the transfer component.

From this brief outline of the basic strategies adopted in MT systems it is evident that all systems require some linguistic analysis of SL texts. It is equally clear that all systems must have methods for generating adequate TL texts, good enough at least for revisers to be able to edit them into more acceptable forms for their clients. Important as TL text generation undoubtedly is, the most crucial component of any MT system will always be the analysis of SL texts. Experience has shown that analysis is the most complex of all stages of translation and that it is the adequacy of the methods for analysing natural language texts which determines whether the system as a whole produces satisfactory results.

It should be noted that this is also true for 'interactive' systems which involve considerable human intervention for the resolution of linguistic ambiguities in the SL text and for the production of TL texts. All MT texts require some editing, simply because no system is perfect. The philosophy of interactive systems is that no computer analysis is at present fully satisfactory (and perhaps never will be) and that difficult problems of linguistic analysis can be circumvented by calling upon human expertise at appropriate moments. However, not all analysis can be left to the human intermediary; otherwise there would be no need for computer assistance at all. It is thus still true to say that the adequacy and acceptability of an interactive MT system is determined in large part by the efficiency of its analytical procedures.

2) Morphological analysis and dictionaries.

The first step of analysis in any MT system is the identification of words in the SL text. This is relatively easy in English and most European languages, since words are separated by spaces in written text. However, it soon became apparent to the earliest MT researchers that recognition of 'orthographic words' was not sufficient. It was obviously wasteful for automatic dictionaries to include every inflected form of a noun or a verb, particularly in languages such as Russian and German. The familiar regularities of noun and verb paradigms encouraged researchers to investigate methods of morphological analysis which would identify stems and endings and thus reduce the size of dictionaries. To give an English example, the words *analyzes*, *analyzed,* and *analyzing* might all be recognised as having the same stem *analyz-* and the common endings *-s*, *-ed*, *-ing.* At the same time, identification of endings was expected to assist the recognition of grammatical categories (word classes or 'parts of speech'), e.g. to continue our example: *-s* indicates a plural noun form or a third person singular present verb form, *-ed* indicates a past verb form, and *-ing* a present participle or adjectival form, etc. In the early word-for-word approaches it was expected that such grammatical information could also aid the processes of word rearrangement in TL text production. Irregular forms, such as the plural noun analyses, were a problem generally dealt with in the main dictionary.

The problems of dictionary size dominated much of early MT research, which is quite understandable in view of the limitations of the computer systems then available. Dictionaries had to be stored on magnetic tapes and searched serially. The most efficient method of dictionary lookup, therefore, entailed preliminary sorting of SL text words alphabetically before matching against dictionary entries. The separation of frequently occurring words into a separate dictionary made good sense, as did the compilation of specialized dictionaries for particular scientific subjects. (Some of these micro-glossaries have been very specialized, capable of dealing only with texts, say, on one narrow field of mathematics.) Problems of idiomatic usage argued for the addition of separate idiom dictionaries. Most later 'indirect' MT systems also have multiple dictionaries, mainly however for reasons related to their linguistic strategies. Thus 'interlingual' systems have one or more dictionaries providing information for SL analysis procedures and one or more for TL synthesis procedures. In 'transfer' systems there are in addition bilingual dictionaries for transfer components.

An obvious point to be made, but one which can be easily forgotten when discussing the complexities of linguistic analysis, is that the success of any MT system is determined above all by the quality and range of its dictionary information. No MT system, however sophisticated its algorithms for analysis and synthesis, will produce good translations if its dictionaries are inadequate.

3) Syntactic analysis.

The first step beyond the basic word-by-word approach is the inclusion of a few rearrangement rules, such as the inversion of

'noun-adjective' to 'adjective-noun', e.g. in French-English translation. In many early MT systems rearrangement rules were often initiated by codes attached to specific dictionary entries. Other more complex differences of syntactic structure were frequently handled by inclusion of phrases in the dictionary, i.e. rather like idiomatic expressions.

An example of such an approach is to be found in the small-scale experiment set up by the Georgetown University research group in cooperation with IBM. The demonstration of this Russian-English translation system in 1954 aroused considerable interest, alerting the public at large to the feasibility of MT of some kind and in part stimulating the flow of U.S. governmental funds into MT research. However, with a vocabulary of just 250 Russian words, only six rules of grammar and a carefully selected sample of Russian sentences, the system demonstrated had little scientific value.

The six grammar rules of the Georgetown-IBM system were as follows (Dostert 1955; Pendergraft 1967; Garvin 1972:51-64):

1) order in SL text to be followed in TL text

2) order in SL text to be inverted in TL text

3) choice of TL form to be determined by indication in following SL word

4) choice of TL form to be determined by indication in preceding SL word

5) the SL word to be omitted, i.e. no TL form

6) insertion of TL form where nothing occurs in the SL text

The rules were initiated by codes attached to dictionary entries; both stems and endings were included in the Russian dictionary since suffixes were identified in the system. Many ad hoc decisions were incorporated, particularly on the insertion and omission of lexical items (rules 5 and 6) in order to deal with 'idiomatic' usages. The restriction of the rearrangement rules to information from the immediate context was accepted solely for the purposes of the demonstration. It was realised that they would be quite insufficient in a larger-scale system. Nevertheless it was believed that the principle operations necessary for MT had been demonstrated.

In many respects, subsequent research at Georgetown was essentially aimed at developing these principles. In 1964 Russian-English systems were delivered to the U.S. Atomic Energy Commission and to Euratom in Italy. The research team adopted what Garvin was later (Garvin 1972) to call the 'brute force' method of tackling problems: a program would be written for a particular SL text corpus, tested on another corpus, amended and improved. The result was a monolithic program of intractable complexity, with no clear separation of those parts concerned with SL analysis and those parts concerned with TL production. Codes for rearrangement were based on morphological information (e.g. type of noun ending), or on conjunction of grammatical categories (e.g. adjective next to noun), or on sentence position (e.g. noun at beginning), or on any other information which could be applied to handle specific situations. The grammatical information in the system was, therefore, contained in codes embedded in the very structure of the program itself and subsequent modification of the system became progressively more and

more difficult (Kay 1973) - in fact both Georgetown systems remained basically unchanged after their installation until the early 1970's.

Although rearrangement rules took into account longer segments of text there was still no notion of grammatical rule or syntactic structure. Syntactic analysis aims to identify three basic types of information about sentence structure:

1) the sequence of grammatical elements, e.g. sequences of word classes: art(icle) + n(oun) + v(erb) + prep(osition) ..., or of functional elements: subject + predicate.

2) the grouping of grammatical elements, e.g. nominal phrases consisting of nouns, articles, adjectives and other modifiers, prepositional phrases consisting of prepositions and nominal phrases, etc. up to the sentence level.

3) the recognition of dependency relations, e.g. the head noun determines the form of its dependent adjectives in inflected languages such as French, German and Russian.

The objectives of syntactic analysis include at least:

a) the resolution of homographs by identifying word classes, e.g. whether watch is a noun or a verb.

b) the identification of sequences or structures which can be handled as units in SL-TL transfer, e.g. nouns and their associated adjectives.

The following descriptions will be partly illustrated by analyses (whole or partial) of the sentence *The gold watch and chain* were *sold by the jeweller to a man with a red beard.* This is a passive sentence (the grammatical subject is the object of the verb), containing a homograph *(watch),* an ambiguous coordinate structure (are both the watch and the chain modified by *gold?)* and three prepositional phrases each of which could in theory modify the verb or their preceding noun phrase.

An example of an analysis program to identify sequential information was the Predictive Syntactic Analyzer developed at Harvard University by Sherry, Kuno and Oettinger (Plath 1967). The premiss was that on the basis of an identified grammatical category (article, adjective, noun, etc.) the following category or sequences of categories could be anticipated with an empirically determinate measure of probability. The system had the following characteristics: under the general control of a push-down store (i.e. last in first out) a sentence was parsed one word at a time left to right, the action taken for each word being determined by a set of predictions associated with the word class to which the word had been assigned in dictionary lookup. At the beginning of the analysis certain sentence types were predicted in terms of word class sequences. Examination of each word was in two stages: first to test whether its class 'fulfilled' one of the predictions, starting from the most probable one, then either to alter existing predictions or to add further predictions. Formally, the system was an implementation of a finite state grammar (fig.4). Initially, only the single most probable path through the series of predictions was taken during parsing, but the results were often unsatisfactory and so later models adopted a multiple-path technique in which all possible predictions were pursued. However, this too proved unsatisfactory: the designers hoped, of course, that multiple parsings would occur if and only if the sentence was genuinely ambiguous. In practice, the analyzer either

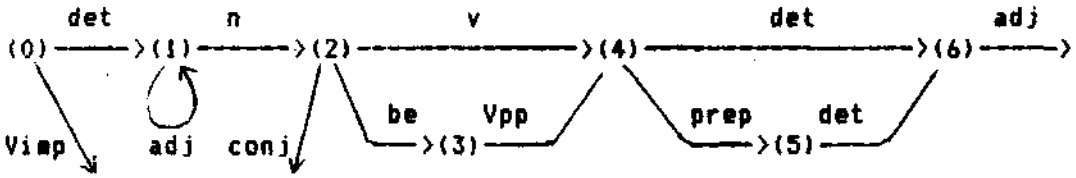


Fig.4. Finite state grammar

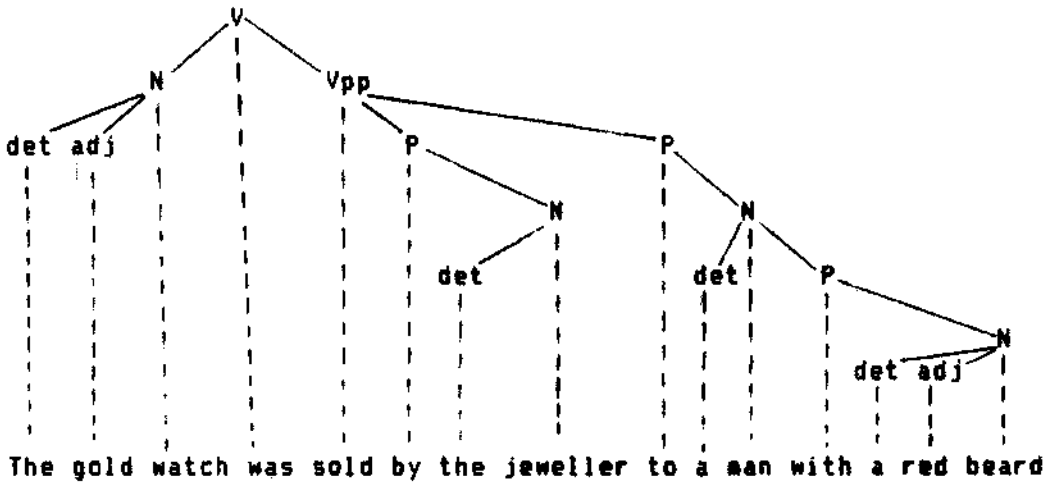


Fig.5. Dependency structure analysis

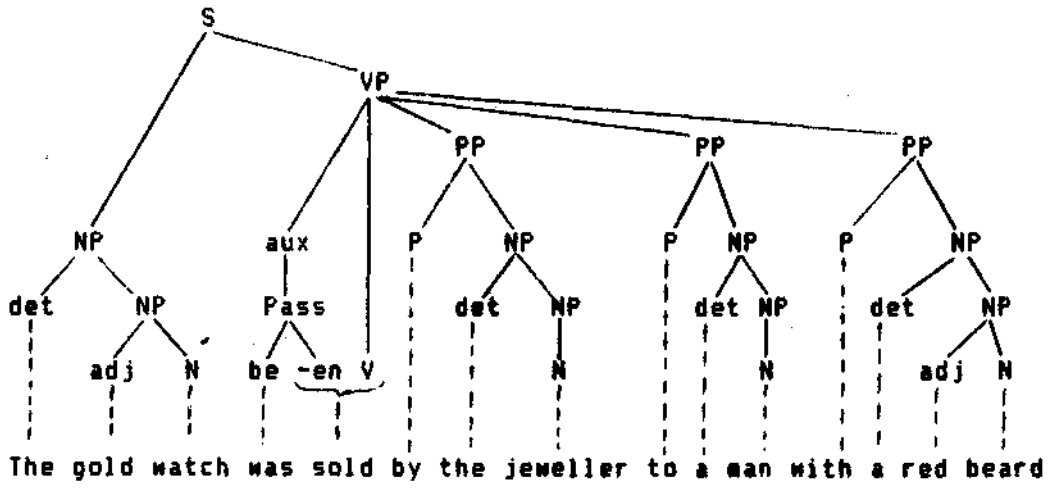


Fig.6. Phrase structure analysis

failed to produce a parsing corresponding to particular acceptable interpretations or produced parsings which corresponded to no acceptable interpretations.

The analysis of dependency relations can be illustrated by the research of Paul Garvin, who had been the member of the Georgetown team principally involved in the 1954 demonstration. He had left the project in 1960 to develop the 'fulcrum' approach to syntactic analysis (Garvin 1972). The basic idea was that in a series of passes the algorithm would identify first the key elements of the sentence, e.g. main finite verb, subject and object nouns, prepositional phrases, then the relationships between sentence components and finally the structure of the sentence as a whole (fig.5). It should be noted that the method was not based purely on syntactic information, it also used statistical information on the probabilities of particular structures given certain configurations of grammatical forms. These probabilities were tested when further information on higher-level relationships became available. Thus the initial identification of a Russian form as a genitive singular noun might be revised in a broader context as a nominative plural noun. The heuristic nature of this approach was deliberately modelled on current AI research on problem-solving systems and, in this respect, foreshadows more recent applications of AI techniques in MT research. Another aspect worth noting also was the proposal to use inter-sentence relationships to determine the subject of sentences. However, the fulcrum method itself proved unsuccessful. After ten years' work at Wayne State University on a Russian-English system a very complex program was still unable to parse Russian sentences with more than one finite verb (Josselson et al. 1972).

The third approach, that of phrase structure analysis (fig.6), was adopted, among many others, by Yngve at MIT (Yngve 1967). As in the fulcrum method, structures were built up in a series of analyses from immediate constituents, e.g. first noun phrases, then prepositonal structures, then verb relationships and finally the sentence structure as a whole. This bottom-up parsing strategy was the most common approach, but at MIT some investigation was made into top-down strategies in which tests are made for a particular structure, say a noun phrase, by checking against word classes. This strategy is now probably more usual than bottom-up. Yngve was also the first to formulate the 'transfer' approach to MT system design in some detail. The essence of his proposal was that the phrase structure analyses of SL sentences should be converted by a transfer component into equivalent phrase structures of the TL (Yngve 1957).

4) Deep and surface structure.

By this time, Chomsky (1957) had demonstrated the inherent inadequacies of finite state grammars, phrase structure grammars and the formally equivalent dependency grammars for the representation and description of the syntax of natural languages. He proposed a transformational-generative model which linked 'surface' phrase structures to 'deep' phrase structures by transformational rules. Thus a passive construction in a 'surface' representation is related to an underlying active construction in a 'deep' representation, where the 'surface' subject noun becomes the 'deep' logical object (fig.7). The Chomskyan model appealed to many researchers. For example, it

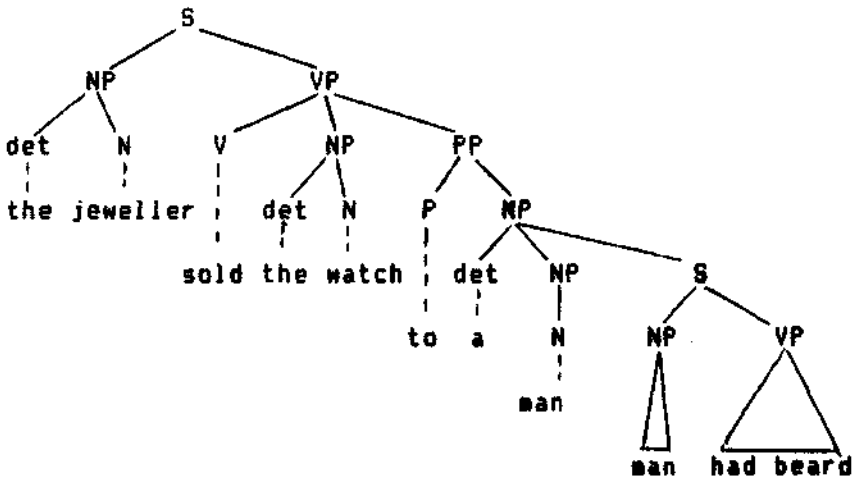


Fig.7. 'Deep' structure analysis

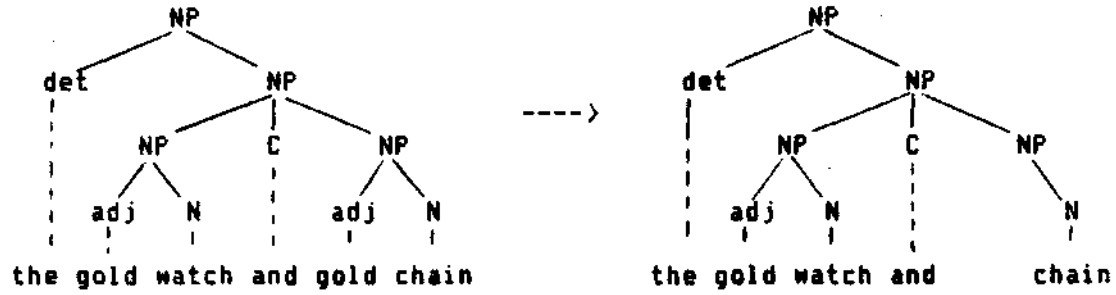


Fig.8. Transformational rule
(loss of phrase structure relationship)

seemed to offer solutions to problems of syntactic ambiguity, such as the relationship of prepositional phrases within sentence structures (cf. figs. 6 and 7). Furthermore, there was the claim that while languages may differ considerably in 'surface' structures they all share the same 'deep structures'. The theory seemed to offer a way of dealing with syntactic equivalencies between languages. Consequently, it proved a stimulus to research on 'interlingual' approaches to MT.

It was soon found, however, that the Chomskyan formulation of transformational rules could not be implemented in a syntactic analysis program. One basic reason for this is that Chomsky's 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. (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 many times argued that MT is a mistaken objective.) Researchers such as Petrick (cf. Grishman 1976) 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 (fig.8) eliminate information from deep structures and there is no way this information can be reconstructed with certainty.

Although transformational rules of the Chomskyan kind are now rarely found in MT systems the distinction between 'surface' and 'deep' syntactic relations pervades nearly all MT projects, and the idea of transforming one structural representation into another has been widely adopted.

As an example of the usefulness of the 'surface' - 'deep' distinction we may cite one of the most successful MT systems, SYSTRAN (Toma 1977a,b; Van Slype & Pigott 1979). In basic conception, SYSTRAN goes back to the Georgetown model. An essential difference is a modular structure which has enabled it to introduce improvements with greater facility. Its parser can be improved by applying new techniques wherever they seem appropriate. Examination of SYSTRAN analysis procedures reveals a mixture of methods.

The first stages of SYSTRAN involve the checking of words of the SL text against a High Frequency dictionary and a Master Stem dictionary. These supply information on grammatical properties (and some semantic features also) which are used in the analysis procedures. Syntactic analysis consists of seven 'passes' through the SL text:

1) the resolution of homographs by examination of the grammatical categories of adjacent words (i.e. sequential information)
2) identification of compound nouns (e.g. *blast furnace*) by checking a Limited Semantics dictionary
3) identification of phrase groups by searching for punctuation

marks, conjunctions, relative pronouns, etc. (i.e. rudimentary phrase structure analysis)

4) recognition of primary syntactic relations such as adjective-noun congruence, noun-verb government and noun-noun apposition (i.e. dependency relations)

5) identification of coordinate structures within phrases, e.g. conjoined adjectives or nouns (i.e. phrase structure)

6) identification of subjects and predicates (i.e. dependency relations)

7) recognition of prepositional structures (i.e. phrase structure)

It is at the stage of recognising subject and predicate relations that SYSTRAN also identifies 'deep' subjects and objects. For example (Billmeier 1982), the passive sentence *The texts were translated by a computer* would be analysed roughly as:

Sentence
- Predicate: verb, past passive.. ..... *translate*
  - Deep subject: ..... *computer*
  - Deep object: ..... *texts*
  - Subject: noun ..... *texts*
    - Determiner: def.art.. ..... *the*
  - Prep. phrase 1: preposition..... *by*
    - Noun phrase: noun. ..... *computer*
      - Determiner: def.art.. ..... *a*

Likewise, a noun phrase containing 'deep' subject-predicate relationships would receive a parallel analysis. Thus, the phrase *the translation of texts by computer* would be analysed roughly as:

Sentence
- :
- (Subject): verbal noun. ..... translation
  - Deep subject: ..... *computer*
  - Deep object: ..... *texts*
  - Determiner: def.art..... *the*
  - Prep. phrase 1: preposition..... *of*
    - Noun phrase: noun ..... *texts*
  - Prep. phrase 2: preposition..... *by*
    - Noun phrase: noun ..... *computer*

The utilization of techniques based on theoretical linguistic models does not, however, change the essentially empirical approach of SYSTRAN - nor its characterisation as a 'direct translation' system. As in its Georgetown ancestor, analysis of SL text goes only as far as the minimum necessary to facilitate reasonable TL translation, and any mixture of types of information is acceptable. Thus, a routine to insert definite and indefinite articles when translating into English from Russian combines syntactic information about the Russian SL text (e.g. whether the noun is qualified by a following genitive noun, a prepositional phrase or a relative clause), some semantic information (e.g. whether the Russian is an ordinal number) and information on English equivalents (e.g. an English 'mass' noun such as *water* usually requires a definite article). Such ad hoc mixtures are common in SYSTRAN and in comparison with later MT systems SYSTRAN lacks a clearly formulated linguistic model. Nevertheless, the undoubted

practical success of SYSTRAN is a strong argument for linguists becoming more familiar with its procedures, particularly since MT systems based on more sophisticated linguistic techniques have not so far any more successful and have in fact in a number of cases proved to be failures.

5) Syntax and semantic problems.

Although the identification of grammatical categories and of sentence structures is clearly important in linguistic analysis, it is equally clear that semantic information is even more crucial for satisfactory translation. The inherent limitations of syntactic information were recognised long before there were efficient parsers. A familiar example is the problem of multiple analyses of prepositional phrases. Since a prepositional phrase may modify either a verb or a preceding noun phrase a sequence such as V + NP_1 + P + NP_2 + P + NP_3 must have parsings which relate NP_2 and V, NP_2 and NP_1, NP_3 and V, NP_3 and NP_2 in all possible combinations. Syntactic analysis alone cannot decide which relationship is correct in a particular case. For example take the sentences:

The coastguard observed the yacht in the harbour with binoculars.
The gold watch was sold by the jeweller to a man with a beard

In the first case, it was the coastguard who had the binoculars; therefore the PP *with the binoculars* modifies the verb. But in the second case, the PP *with a beard* modifies the preceding noun *man.* Only semantic information can assist the analysis by assigning semantic codes allowing binoculars as 'instruments' to be associated with 'perceptual' verbs such as *observe* but prohibiting *beards* to be associated with objects of verbs such as *sell.*

Such solutions have often been applied in MT systems. However, they cannot deal with all problems of syntactical ambiguity. As Bar-Hillel argued in 1960 (Bar-Hillel 1964), human translators frequently use background knowledge to resolve syntactical ambiguities. His example was the phrase *slow neutrons and protons.* Whether *slow* modifies *protons* as well as *neutrons* can be decided only with subject knowledge of the physics involved. Similarly, in the case of the *gold watch and chain* our assumption that both objects are gold is based on past experience. On the other hand, in the case of the phrase *old men and* women the decision would probably rest on information conveyed in previous or following sentences in the particular text being analysed. The incorporation of such information in an automatic parser leads to obvious exponential complications of the procedures.

Syntactic ambiguity is, of course, only one aspect of ambiguity in language analysis. Words of more than one meaning are a perpetual problem. It is true that homographs such as *watch* can often be distinguished by syntactic analysis alone, i.e. whether the word is a noun or a verb. However, the resolution of some ambiguous words require, as in the physics example, knowledge of the objects referred to. There is, for example, a third sense of *watch* in the sentence: *The watch included two new recruits that night.* It can be distinguished from the other noun only by recognition that time-pieces do not

usually include animate beings. In an influential paper, Bar-Hi11el (1960) argued that fully automatic translation of a high quality was never going to be feasible. (His example was the sentence *The box was in the pen* where the elimination of the writing implement sense of *pen* can be achieved only with knowledge of the relative sizes of *box* and *pen* in the particular context of the text being analysed. It is true that in practice this type of problem can be lessened if analysis is restricted to a more or less narrow scientific field, and so dictionaries and grammars can concentrate on a specific 'sublanguage' (Kittredge & Lehrberger 1982). Nevertheless, similar examples recur regularly in discussions on the inherent feasibility of MT. Consequently, the argument that fully automatic translation presupposes 'language understanding' based on encyclopaedic knowledge and complicated inference procedures has convinced many researchers that the only way forward is the development of 'interactive' MT systems. For others, it has been a powerful stimulus to experiment with systems incorporating some of the techniques and methods of artificial intelligence.

It should be clear from these comments that semantic analysis has developed, by and large, as an adjunct of syntactic analysis. As we have seen, in many MT systems semantic analysis goes no further than necessary for the resolution of homographs. In such cases, all that is generally needed is the assignment of such features as 'human', 'animate', 'concrete', 'male', etc. and some simple feature matching procedures. For example, *crook* can only be animate in *The crook escaped from the police,* because the verb *escape* demands an animate subject noun. The 'shepherd's staff' sense of *crook* is thus excluded. In 'direct translation' systems such as SYSTRAN this is usually the limit of their semantic analysis. A consequence, however, is that semantic features are often assigned ad hoc, as the demands of the SL analysis and the TL text production seem to require them. The absence of systematic generalization in the application of semantic features is perhaps, as Pigott (1979) has shown, one of the principal deficiencies in the SYSTRAN English-French translation system. Furthermore, semantic features tend to be applied rather rigidly. There are difficulties, for example, if the verb *sell* is defined as always having inanimate objects; the sentence *The men were sold at a slave market* would not be correctly parsed.

6) Semantic analysis.

The adoption of 'interlingual' strategies and the need to provide universal 'deep' structure representations stimulated the investigation of more detailed semantic procedures. First, relationships between lexical items were analysed in semantic terms, e.g. in the logical relations of predicates, arguments, entities and attributes, and in the valency (or case) relations of 'agent', 'instrument', 'location', etc. Thus in the sentence *The watch was sold by the jeweller to a man with a red beard* the predicate would be the verb *sell* and its arguments (or dependents) *watch, jeweller* and *man*, and the latter would be respectively 'object', 'agent' and 'benefactive' of the transaction. Secondly, following the theoretical speculations of linguists researchers investigated the possibilities of analysing lexical items into semantic features common to all

languages, i.e. universal semantic primitives. For example, *boy* would have the features 'human', 'male', 'young'; and *kill* the features 'cause', 'become' and 'die' in a suitable relationship.

Initially, the first of these aspects was the subject of most activity in the development of 'interlingual' MT systems. An example is the German-English translation system METAL developed at the University of Texas in the 1960's and 1970's (Lehmann & Stachowitz 1972-75); the later METAL system funded by Siemens since 1980 has taken a more 'transfer'-1ike approach. The aim was to develop analysis procedures and forms of universal representations which could be easily adapted to other pairs of languages. Analysis was performed by three 'grammars' working in sequence. After morphological analysis and dictionary lookup, the first two stages produced tentative 'standard strings', i.e. possible sequences of grammatical categories, and then tentative 'standard trees', i.e. phrase structures. The third stage, 'normalization', filtered out the semantically ill-formed phrase structures by reference to the semantic features provided by dictionary entries and then converted each acceptable 'standard tree' into a 'normal form' (or several 'normal forms' if the sentence was genuinely ambiguous). In this 'deep structure' representation relationships between items were expressed in terms of 'predicates' and 'arguments' or, alternatively, 'entities' and 'attributes'. A sentence such as *The old man in a green suit looked at Mary's dog* would receive the 'normal form' in fig.9. This conversion procedure involves the identification of *in* as the predicate element of a tree *with the green suit* and *the old man* as argument elements; and the recognition of the adjectives in these noun phrases as arguments of their respective head nouns. The result is a dependency-style semantic representation intended to be independent of language-specific surface and phrase structure forms. It was not, however, a fully interlingual representation. In METAL lexical items were not broken down completely into semantic primitives and so certain types of paraphrase relations could not be handled. For example, *He ignored her* and *He took no notice of her* would not be recognised as equivalents because these sentences would have different 'deep' structures.

A similar philosophy inspired the CETA approach at Grenoble University, a project which also began in the early 1960's (Vauquois 1975). As in METAL, the first stages of analysis were morphological analysis and dictionary lookup, followed by a phrase structure analysis. The next stage converted these 'surface' structures into dependency trees of a more explicit and rather more abstract form than those in METAL. Fig.10 illustrates the semantic representation of *The formula explains the frequent occurrence of neutrons.* It is derived from a phrase structure, first, by the addition of dependency relations such as the marking of a verb as 'governor' and of a noun phrase as 'dependent'; then, by the classification of lexical items as either predicatives or non-predicatives, the former including adjectives and adverbs as well as verbs and the latter including nouns and articles; and finally, by analysing the whole structure in terms of predicative and arguments ('actants').

The Grenoble approach to MT system design has been influenced most by the stratificational model of the Russian linguist Mel'chuk

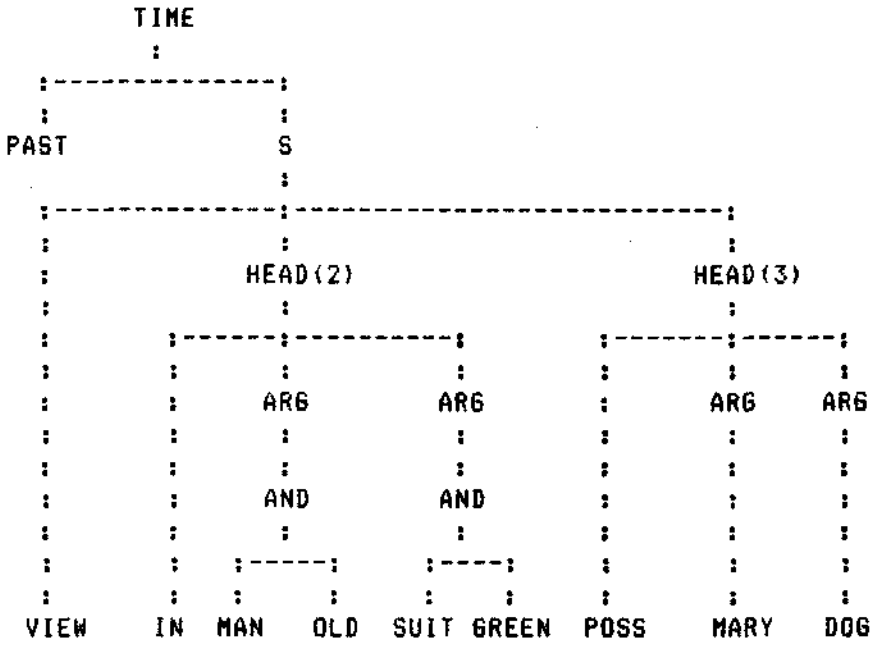


Fig.9. METAL representation

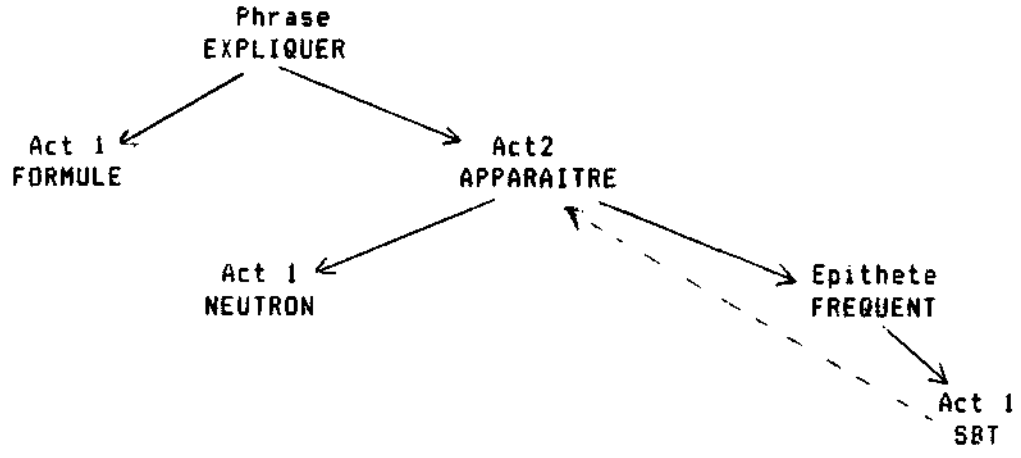


Fig.10. CETA representation

(Mel'chuk & Zholkovskii 1970). Like the analogous but nevertheless distinct and independent stratificational theory of Lamb (1966), Mel'chuk's original conception developed from work in MT (e.g. Kulagina et al. 1971), but it has remained more firmly rooted to the practicalities of MT analysis than Lamb's more theoretical speculations. In Mel'chuk's 'meaning-text' model there are five basic levels or 'strata' of linguistic representations: phonemic, morphological, surface syntactic, deep syntactic, and semantic. Surface syntactic representations include such grammatical dependency relations as 'subject-of', 'complement-of', 'auxiliary' and 'determinant'; deep syntactic representations include valency relations such as 'agent', 'instrument' and 'location'; and semantic representations are abstract networks of semantic primitives.

From the description of the CETA system it is evident that it also was not a fully interlingual system. It is true that CETA could deal with some syntactic equivalences, e.g. the structure dominated by *apparaître* in fig. 10 could represent either a subordinate clause with *apparaître* as finite verb or a noun phrase with *apparition* as a verbal noun. Nevertheless, it lacked much of the detailed paraphrasing operations present in Mel'chuk's model, which result from the indication of complex semantic relations. These include, as one would expect, such relations among individual lexical items as synonymy, antonymy and conversives, e.g. *fear* and *frighten,* a verb and an agentive noun, e.g *write* and *writer*, *prevent* and *obstacle,* or a verb and its causative form *lie* and *lay*. They include also phraseological and idiomatic constructions, such as indications of the typical or 'idiomatic' verb for expressing particular relations to a given noun, e.g. the inceptive verb for *conference* is *open* but for *war* it is *break out*, the causative verb for *dictionary* is *compile*, for *foundations* it is *lay* and for a *camp* it is *set up* or *pitch*, and the realisational or implementative verb for *order* is *fulfill,* for *law* it is *observe,* for *promise* it is *keep* and for *obligations* it is *discharge.*

Mel'chuk has not by any means been the only MT researcher to tackle the semantic complexities of natural language. The treatment of phraseological relations has much affinity to Ceccato's approach to interlingual semantic structures, which he called correlational analysis (Ceccato 1967); and it recalls also aspects of the 'thesaurus' approach to MT semantics of the Cambridge Language Research Unit (Masterman 1957). At the present time, the most obvious affinity is with the work on semantic representation in the context of research on 'text understanding' in artificial intelligence.

7) Artificial intelligence methods.

Linguistic analysis in AI research has not been directly concerned with translation problems but rather with question-answering systems and general problems of language understanding. The relevance of AI research to problems of semantics in MT systems has been obvious to many researchers, and AI techniques are applied increasingly in translation systems. Nevertheless there have been few explicitly AI projects on MT as such.

Characteristic of the AI approach to linguistic analysis is the abandonment of syntax-based models in favour of predominantly

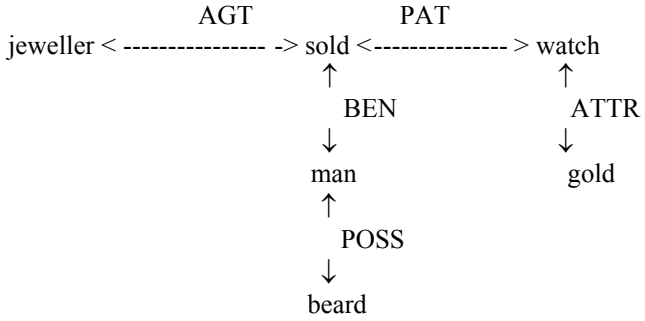


Fig.11. Semantic representation (Wilks)

```
SCLAB3:
   SCRIPTNAME $VEACCIDENT
   MAINCON EVNT4
   SCENECONS (EVNT4 EVNT17 EVNT33)
   INFERENCE ((EVNT20 EVNT26) (EVNT14))
   SCORECARD (EVNT33)

EVNT4:
   VALUE ((ACTOR STRUCTO
                <=> (*PROPEL*)
                OBJECT PHYSO)
          TIME (TIME))
   LASTEVENT (EVNT3)
   NEXTEVENT (EVNT20 EVNT14 EVNT7)

STRUCTO
   CLASS (£STRUCTURE)
   TYPE (*CAR*)
   SUPERSET (*VEHICLE*)
   ELEX (AUTOMOBILE)
   SLEX (AUTO)
   SROLES
       ((*VEHACCIDENT . &VEHICLE)
       ($DRIVE . &VEHICLE1))

PHYSO
   CLASS (£PHYSOBJ)
   TYPE (*TREE*)
   ELEX (TREE)
   SLEX (ARBOL)
   SROLES
      (($VEHACCIDENT . &OBSTRUCTION) )

EVNT14:
   VALUE ((ACTOR HUMO
                FROM (*HEALTH* VAL (*NORM*)))
                TOWARD (*HEALTH* VAL (-10)))
                TIME (TIME17))
```

Fig.12. Semantic representation (Carbonell)

semantics-based models. Semantic analysis is not seen as just the next stage after syntactic analysis, in effect to tackle problems remaining after syntactic analysis, but as the central component of the system. Problems of syntactic structure are left, if necessary, to subsidiary operations.

One of the first to experiment with an AI semantics-based approach was Wilks in his prototype English-French MT system in the early 1970's (Wilks 1973, 1975a). The SL text is first partitioned at punctuation marks and 'function words' (prepositions and conjunctions) into fragments, e.g. *I advised him to go* becomes '(I advised him) (to go). Each fragment is then tested against an inventory of templates, triples of semantic features. For example, the template MAN HAVE THING (paraphrased perhaps as "some human being possesses some object") would be matched on a sentence such as *John owns a car*. MAN, HAVE and THING are intended to be interlingual semantic primitives which would be found as the principal semantic features of the words *John, own* and *car* respectively. Semantic formulas or definitions of words are constructed from semantic primitives, e.g. the formula for *drink* is: (*ANI SUBJ) (((FLOW STUFF) OBJE) ((*ANI IN) (((THIS (*ANI (THRU PART)))TO)(BE CAUSE))))). This is to be read as "an action, preferably done by animate things (*ANI SUBJ) to liquids ((FLOW STUFF)OBJE), of causing the liquid to be in the animate thing (*ANI IN) and via (TO indicating the direction case) a particular aperture of the animate thing; the mouth of course" (Wilks 1973) The semantic analysis of lexical items goes no further than necessary; in this context there is no need to distinguish *mouth* from other apertures. The notion of preference is a central feature of Wilks' method: SUBJ displays the preferred agents of actions and OBJE the preferred objects or patients, they do not stipulate obligatory features of agents and patients, and thus allowance is made for 'abnormal' usages, e.g. *The car drinks petrol* and *The men were sold in a slave market.* In this way, Wilks' preference semantics can cope with many types of metaphorical expressions without adding to the complexities of dictionary entries (Wilks 1975b). The final stage of the analysis produces a dependency network of semantic relations on the basis of the valency (or case) links specified. Thus, our example sentence *The watch was sold by the jeweller to a man with a beard* might receive the analysis in fig. 11.

At this point relationships between the networks of fragments are established; thus a temporal phrase (*during the war*) might be tied to the 'action' element of an earlier or later fragment. It should be noted that ties are made not only within sentences but also across sentence boundaries, since the basic unit is not the sentence but the phrase (fragment). Some ties involving pronominal reference make use of 'common sense inference' rules. For example, in the sentence *The soldiers fired at the women and we saw several of them fall* the linking of the pronoun *them* to the noun *women* rather than to the other noun *soldiers* is made on the basis of a common sense rule stating that if an animate object is hit then it is likely to fall. In other words this rule establishes a causal relationship between the components of the templates of the fragments in question.

A more advanced mechanism in AI for the making of inferences is embodied in the notion of 'scripts'. At Yale University, a

rudimentary interlingual MT system has been developed by Carbonell and others (1978) using the 'story-understander' model of the Schank AI team. A simple English text, the report of an accident, is analysed into a language independent conceptual representation by referring to 'scripts' about what happens in car accidents, ambulances, hospitals, etc. The resulting representation is the basis for generating Russian and Spanish versions of the original report.

An extract of the semantic representation is given in fig. 12 for the sentences:

Friday evening a car swerved off Route 69. *The vehicle struck a tree. The passenger ... was killed…*

The main structure (SCLAB3) indicates the type of script <$VEHACCIDENT), the main story in EVNT4, episodes of the story (SCENECONS) and the inferred events or actions (INFERENCE). The representation EVNT4 expresses the crash itself: a 'structured physical object' (STRUCTO) hitting (*PROPEL*) an 'unstructured physical object' (PHYSO). The representations of STRUCTO and PHYSO indicate that these are respectively a *VEHICLE*, specifically an AUTOMOBILE (*a car*), and a *TREE* (*a tree*). Connected to EVNT4 are other events, including the result (EVNT14) that someone (HUMO) died, i.e. went from state of normal health to state of non-health.

These examples of the AI approach to MT give some flavour of the complexity demanded in semantics-based methods. Their feasibility in full-scale MT systems must remain in some doubt for the present. The problems for a large-scale MT system are clear enough: the establishment of semantic primitives to be applied methodically in the creation of large dictionaries; the establishment of templates or scripts to cover a much wider range of possible texts; and the establishment of 'inference' rules and large knowledge databases to handle all the possible relationships which may occur in the texts to be translated. AI workers such as Carbonell are confident that there is no evidence of a combinatorial 'explosion' when systems are expanded to cover wider ranges of texts. These assurances are treated with understandable scepticism by MT workers in view of past experience with the problems of large systems.

8) Pragmatic aspects.

However, there are further more fundamental reservations. Experience with 'interlingual' approaches such as CETA, which like AI systems attempt to derive language-independent representations, demonstrated that the analysis of SL texts to such levels of abstractness entails the loss of information which can be valuable if not essential for TL synthesis. In the transition from the 'surface structure' representation to the 'deep' structure was lost about which noun (or NP) was the grammatical subject, which the object, whether a passive construction was used in the SL text, which was the main clause and which the subordinate clauses. In other words some of the sequential information has been lost. Yet this information can be vital for the generation of TL equivalents. In Mel'chuk's model such information is retained, but in CETA it was not. It is also obvious that this is true also in Carbonell's prototype system: the conceptual representation is explicitly intended not to reflect the surface form of the input text, it is a 'summary' paraphrase of the content;

consequently the TL output cannot be a true translation.

The other fundamental lesson from experience with 'interlingual' MT was that systems based on a 'filtering' conception of analysis could be too rigid. The failure of any one of the analytical procedures to produce an acceptable representation meant that the whole process failed. Morphological analysis could fail because the dictionary had no entry for a particular word or because it did nor record all homographic variants. Syntactic analysis could fail if it could not parse any part (however small) of a sentence.

Since the mid 1970's MT research has tended to favour the somewhat less ambitious 'transfer' approach to system design. It is no longer the aim to analyse SL texts into interlingual representations but to analyse to a depth sufficient for effective transfer into TL deep representations. Such transfer or 'interface' representations may include a mixture of language-specific and language-universal semantic features, they may also include a mixture of 'surface' and 'deep' syntactic information, they may include dependency information or phrase structure information. There is a greater flexibility in the 'transfer' strategy in comparison with earlier approaches. The flexibility is to be found not only in a less rigid hierarchy of levels but in the use of more powerful and more generalised techniques of analysis.

9) General analysis methods.

Current MT research favours two general approaches to analysis techniques. One is derived from the work of Woods on augmented transition network parsers; the other is what may be called the 'tree transduction' approach, deriving in part from the Q-systems developed in the TAUM project. These are not, of course, the only parsers available; in recent years, research in computational linguistics has produced many efficient analysis programs, e.g. the one developed by the Linguistic String Project (Sager 1981) and successfully applied in information retrieval systems. (A good survey of analysis programs is Grishman 1976.)

The parser developed by Woods (1970, 1973) consists of a series of finite state transition networks. Woods' parser differs in two important respects from the finite state 'grammar' of the Predictive Syntactic Analyzer already mentioned (fig.4). Firstly, the arcs of one network may be labelled with the names of other networks; thus, in the extremely simple 'grammar' of three networks displayed in fig.13, 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)', 'adj(ective), 'n(oun)' as necessary 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. Its 'transformational' capability is achieved by adding tests and conditions to the arcs and by specifying 'building

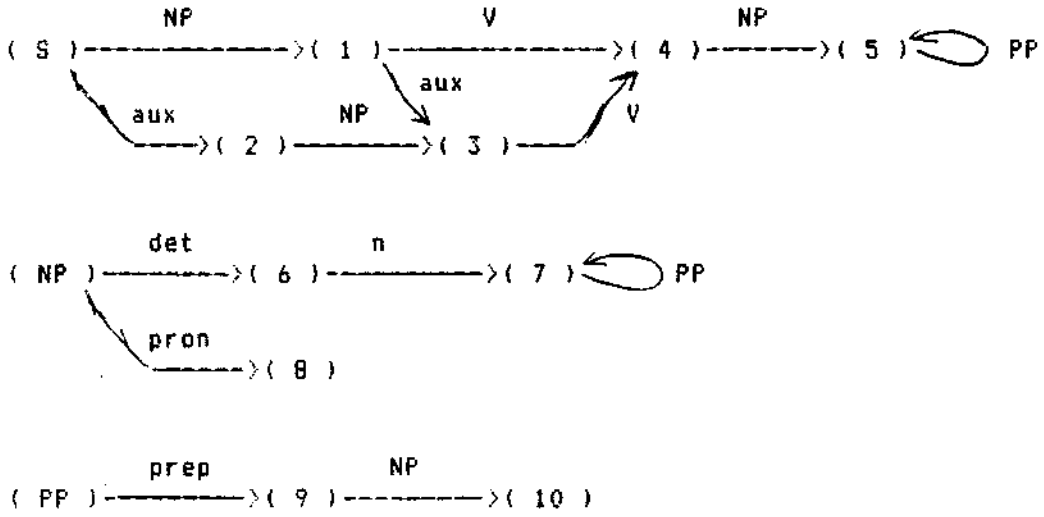


Fig.13 Partial ATN grammar

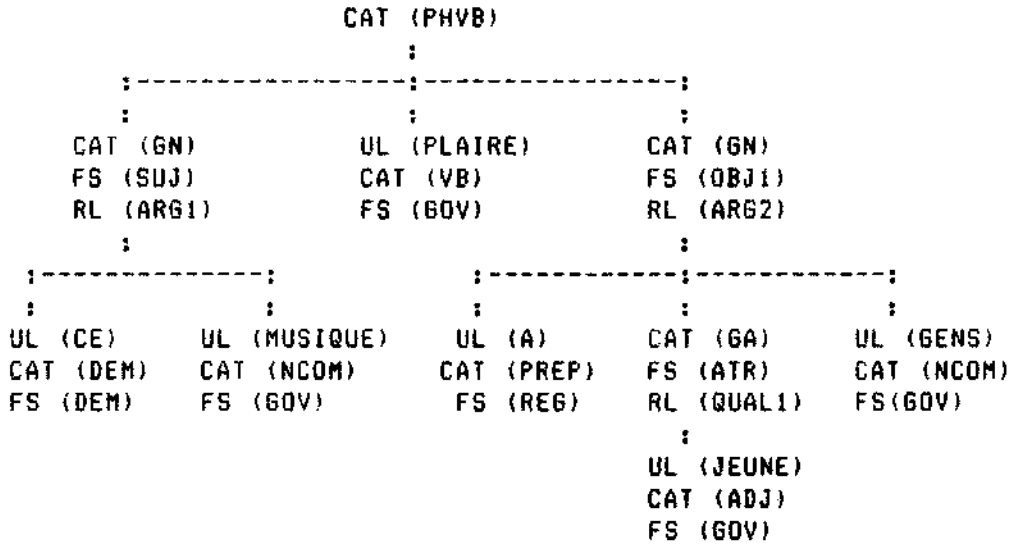


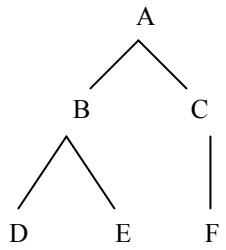
Fig.14. GETA-ARIANE representation

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.

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.

Its principal disadvantages are those common to top-down parsers. Lower level constituents have to be analysed every time and always in same way whenever a particular higher level structure is being tested (e.g. the NP in a subordinated prepositional phrase). In complex sentences this redundancy can damage efficiency. Another problem for ATN parsers is the treatment of coordinate constructions. Since coordination can involve almost any level of analysis, e.g. single noun, noun phrase, clause and sentence, no prediction can be made about the level at which a particular conjunction is operating. The parser cannot, therefore, predict the scope of a conjoined structure and know exactly when to return to the basic structure.

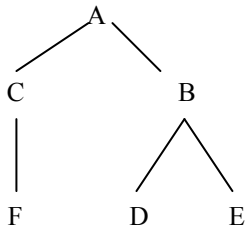
Like ATN parsers, algorithms for the transformation of one tree structure into another can be applied at many stages of linguistic analysis. Such algorithms are based on the fact that any tree can be expressed as a string of bracketed elements, thus the tree:



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:



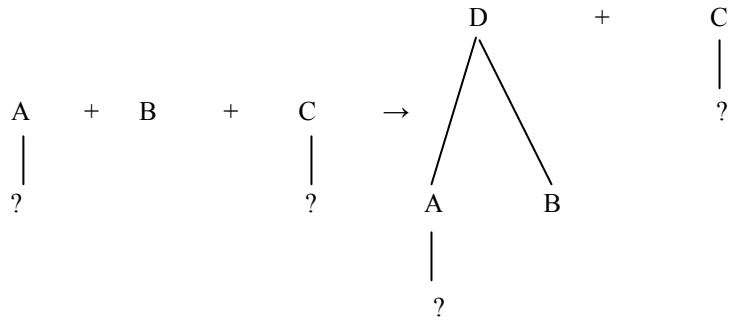
Tree-transducers are able to deal with the occurrence of optional elements in trees or sub-trees which are not affected by the conversion rules, e.g. the occurrence of an unspecified string or tree '?' between B and C at the same level. (For example, B and C might be elements of a phrasal verb *look...up.)* 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. An example of the latter is morphological analysis, illustrated by the use of the TAUM project's Q-system (TAUM 1973). First, the system deals with irregular forms and the 'function words'. A small dictionary is consulted to assign category labels to prepositions, conjunctions, articles, pronouns, and also some idiomatic prepositional phrases. Thus, the strings WITHIN, SEVERAL, and IN + THE + PROCESS + OF become the subtrees P(WITHIN), QUANT(SEVERAL) and P(INTHEPROCESSOF). In the next stage, prefixes are detached: e.g. UNDERSTOOD becomes UNDER + STOOD; the irregular form STOOD is now regularized as the subtree sequence: SW(STAND) + ED(PST); and then after restoration of the prefix: UNDER + SW(STAND) becomes SW(UNDERSTAND) and the correct stem form can be matched against the dictionary.

i.e.

1) UNDERSTOOD
2) UNDER + STOOD

3) UNDER + SW + ED
STAND PST

4)

The remaining forms are treated as regular constructions. First, regular suffixes are split off: e.g.

ANALYZED → ANALYZ + -ED
TRIED → TRI + -ED
PUTTING → PUTT + -ING
SERIES → SERIE + -S.

Some of these divisions like the last will prove abortive. Then rules for the regular derivation of stem forms are applied, e.g. adding a final E: ANALYZ → ANALYZE
undoing the doubling of final consonants: PUTT → PUT
replacing I by Y: TRI → TRY

These hypothesised stems are then checked against the dictionary, which would assign grammatical categories for the next stage of analysis. Because some incorrect divisions are made the TAUM dictionary lookup had to include the unsegmented forms as well: thus SERIES as well as SERIE. This necessity introduced obvious redundancy, but it was offset in part by the power of the algorithm.

As this example illustrates, tree-transducers like TAUM's Q-system parser are bottom-up parsers. When applied in syntactic analysis, therefore, they build upwards from grammatical categories to phrase structures (NP or VP) to sentence structures. In fact most of the phrase structure and dependency parsers used in MT systems have been bottom-up parsers. The usual problem with this approach is that information obtained at a lower level has to be transferred through each higher level until it reaches the stage where it can be applied.

However, many of the problems which this could cause, e.g. with pronoun-antecedent relations in subordinate clauses, can be overcome by exploiting the powerful flexibility of tree-transducing algorithms. Just as the 'building' instructions in ATN parsers can incorporate any kind of syntactic relation and any kind of semantic representation, so also the labels on tree representations can represent any syntactic or semantic relation. Likewise, the rules for the transformation of strings, trees and subtrees can be constrained by conditions referring to any kind of linguistic data in much the same way as conditions are applied to arc transitions in ATN parsers.

An example is to be seen in the GETA-ARIANE system (Boitet & Nedobejkine 1981), a 'transfer' system developed at Grenoble University as a successor of the 'interlingual' CETA system. Experience with CETA had shown the value of including various levels of grammatical and semantic information in SL representations. In GETA 'deep structure' or SL transfer representations include a mixture of levels of interpretation: syntactic classes (adj, noun, NP, VP) or grammatical functions (subject, object, etc.) or logico-semantic relations (predicates and arguments). In other words, they combine information about phrase structure relations, dependency relations and semantic or logical relations. For example, the sentence:

Cette musique plaît aux jeunes gens

would have the tree representation in fig. 14 where:

a) UL indicates lexical items (MUSIQUE), (GENS) etc.
b) CAT indicates grammatical categories such as noun phrase (GN) and adjective (ADJ)
c) FS indicates dependency relations such governing node (GOV), subject (SUJ) and attribute (ATR)
d) RL indicates logico-semantic relations such as ARG1 and ARG2.

The considerable merit of tree transducers is that their very abstractness and flexibility allow the linguist to experiment with different 'grammars' and types of representations. He can decide what transformations to use in particular instances and what conditions are to be attached to their use; he can construct 'subgrammars' to be applied in any order and under any conditions he may specify. He might, for example, construct a set of different subgrammars for the treatment of noun groups, one for simple cases, another for complex cases. He might apply a strategy using dependency relations in one subgrammar and a strategy using phrase structures in another. In addition, the linguist can be reasonably sure that, whatever the strategy or 'grammar' used, there will always be a result at the end of a finite application of rules. This is because tree-transducing algorithms do not test for the 'acceptability' of structures (i.e. they do not filter out ill-formed structures) but test for the 'applicability' of transduction rules. If a rule does not apply the tree remains unchanged; if no rule of a subgrammar can be applied there will always be a tree as output on which other subgrammars may operate.

However, for some procedures tree-transducers have been found to be too generalised and unnecessarily powerful. In the GETA system, for example, while syntactic and semantic analysis uses a tree transducer (ROBRA) morphological analysis (ATEF) is based on a simple finite state parser. In the TAUM project the Q-system algorithm was replaced by an ATN parser (REZO) for syntactic analysis in the large-scale system for the translation of aircraft maintenance manuals (Isabelle et al. 1978). Tree transducers remained in use for morphological analysis, structural transfer and syntactic synthesis processes for which they are clearly well suited - and in the highly restricted environment of the METEO system for translating English weather reports into French (Thouin 1982), where excessive power was a less critical factor.

10) Flexibility of analysis procedures.

One of the most characteristic features of recent MT system design is this flexibility of approach. Earlier systems tended to implement one particular theory of analysis (e.g. phrase structure). More recent systems have been deliberately designed to test a multiplicity of analytical strategies.

It must be admitted that there could be dangers in allowing too much latitude. The overall perspicuity of the analysis structure might be lost, it might be difficult to maintain consistency in linguistic procedures, and there could consequently be a degree of ad hocness perhaps almost as great as in the early MT systems. To counteract this danger present systems incorporate a modular structure in which the various parts of the analysis processes are kept as independent as possible of each other. As we have seen, the 'direct

translation' system SYSTRAN has a modular structure in its analysis procedures. An example of a 'transfer' system is the Saarbrücken system (SUSY) which has the following analysis modules (Luckhardt 1982; Eggers 1981):

1) text input
2) dictionary lookup, compounds and derivational analysis
3) disambiguation of lexical homographs
4) isolation of sentence segments (phrases and clauses)
5) analysis of nominal phrases
6) analysis of verbal phrases
7) identification of valency relations, complementation and subordination
8) semantic disambiguation

Each of these modules may be further subdivided. For example, in the SUSY Russian program (Haas 1980) the analysis of nominal phrases has the following sub-modules:

a) identification of noun groups on the basis of acceptable sequences of word classes
b) test of adjectival and case agreements
c) identification of attributive relations
d) identification of coordinate relations

Associated with each module is a series of dictionaries or grammatical data. For example, the third module has access to information about the distribution of various word classes, e.g. that in German if *um*, *ohne* or *anstatt* are followed in the same sentence by *zu* then they are not prepositions introducing prepositional phrases but conjunctions introducing infinitive constructions. The sixth module has access to data about verbal phrases, e.g. what kinds of clause they may occur in and whether they are separable or not (cf. *unterbringen* and *unterrichten* or in English *look up* and *look after*) - such information may not only assist in the analysis of sentence structure but also distinguish homographs (e.g. *unterhalten* is separable in the sense 'hold under' and inseparable in the sense 'support').

This kind of modular structure has clear advantages for the construction and testing of programs, for the establishment of banks of linguistic data, for the monitoring of analysis processes, and for the improvement of system design. In particular, it opens up the possibility of cooperative research projects as in the EUROTRA system funded by the Commission of the European Communities (King 1982). Collaborating teams from different countries of the Community are developing analysis and generation components for their respective languages to be brought together eventually in a large-scale multi-lingual 'transfer' system. There are of course some constraints, particularly with respect to the basic structure of interface representations, but each team is free to apply any techniques of analysis they consider most appropriate.

11) Some lessons from past 'failures'.

In the past MT projects have worked more or less independently of each other in the sense that programs and dictionaries produced by one team have not been utilized by others - a deplorable waste of much effort and (mainly public) money - this is still true in many cases. The main explanation is that projects have

been set up to test one particular approach or theory and the work of previous MT research is considered unusable. Naturally, each new theory has been adopted because it promised to solve the difficulties of earlier methods. Thus, phrase structure and dependency analyses were advances on word-for-word systems. Transformational grammars, it was hoped, would overcome the inadequacies of phrase structure models. Semantic features were introduced to solve problems of syntactic ambiguity. Semantic primitives and interlingual representations were proposed as the only real solution to numerous problems of synonymy, ambiguity and semantic incongruities between languages. Finally, the AI approach is seen by its advocates as the only possible answer to the semantic problems of translation.

New techniques have often been advocated with great enthusiasm; researchers have been encouraged by initial results with experimental algorithms and trial dictionaries to predict imminent success. But the expansion and elaboration of methods and techniques to the larger scale and size necessary for practical systems have often revealed insurmountable difficulties. There has been an unfortunate tendency in MT and, as Dreyfus (1972) has illustrated, in AI research also, to succumb to the temptation of predicting 'breakthroughs' on the basis of small-scale initial success. It is frequently overlooked that what determines the success of an analysis procedure, in MT as in other applications, is more often the quality of the linguistic information, the richness and accuracy of the dictionary, than the sophistication of the linguistic or AI theoretical model.

It could well be argued that in the past MT research has been misled by the claims of theoretical linguistics. It has to be admitted that the most successful operational system, SYSTRAN, owes very little to linguistic theory of the past twenty years, while more linguistically advanced systems have been abandoned or have yet to move outside the laboratory. What is the reason for this 'failure' of linguistics to suggest models appropriate to MT? In part it is attributable to the concentration on formal definitions of language systems and the neglect of investigations of language behaviour in social settings; linguistic theory has pursued the goal of scientific rigour, idealisation and abstraction without checking its hypotheses and theoretical models against empirical observations of actual linguistic usage. Paradoxically, therefore, the very impetus for the formalisation of grammars which made the automation of linguistic processes appear a feasible objective has itself encouraged the dissociation of theory and practical reality and to the adoption of unrealisable models. The activity within the AI context on language understanding has corrected a good deal of this excessive abstraction; AI researchers are disinclined to wait until a theory is complete but prefer to experiment with semi-formulated hypotheses. Furthermore, AI has been far more concerned with semantic problems and with tackling the problems of text structure, but of course an operational AI system for translation has yet to appear.

One response to the failures of linguistically sophisticated systems has been the development of interactive systems. The argument is that those problems of linguistic analysis which cannot be tackled by the computer should be left to human intermediaries. The

availability of on-line interrogation facilities and the accessibility of word processors have made possible such interactive systems as the LOGOS, ALPS and Weidner systems. Although interactive systems make use of many of the techniques and methods described they frequently do not attempt to resolve homographs or identify structural relationships such as those of prepositional phrases. These are left to the human translator. The expectation in many cases is that eventually satisfactory methods will be developed elsewhere, perhaps within the AI context, which can be integrated into the working system. In other words, it is essentially the empirical 'boot-strap' approach of earlier MT researchers.

The response of those pursuing the longer term goal of fully (or nearly fully) automatic systems is to take a pragmatic attitude to linguistic and AI theory, to make use of what is appropriate and to develop their own methods where necessary. There is no longer the overreaching ambition to construct complete interlingual systems; there is no longer the obstinate pursuit of a single theory or approach; the present emphasis is on flexibility. The success of SYSTRAN has demonstrated that MT systems should take from linguistics and from AI only those techniques which are appropriate and justifiable. While there are inherent limitations to such flexibility in the 'direct translation' design and strategy of SYSTRAN, the modular, multi-level 'transfer' approach of systems such as SUSY, EUROTRA and GETA promises much greater adaptability. The difficulties of constructing flexible, integrated, coherent, clearly structured, linguistically sound and efficient systems are indeed immense, but there is sufficient experience in the MT community to encourage the expectation that the goal can be achieved within the not too distant future.

REFERENCES

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

Bar-Hillel, Y. (1964)
'Four lectures on algebraic linguistics and machine translation', in his: *Language and information* (Reading, Mass., Addison-Wesley, 1964), 185-218

Billmeier, R. (1982)
'Zu den 1inguistischen Grundlagen von SYSTRAN' *Multilingua* 1, 1982, 83-96

Boitet, C. & Nedobejkine, N. (1981)
'Recent developments in Russian-French machine translation at Grenoble' *Linguistics* 19, 1981, 199-271

Carbonell, J. et al. (1978)
Knowledge-based machine translation. Research report, Yale University, 1978

Ceccato, S, (1967)
'Correlational analysis and mechanical translation', in: Booth, A.D. (ed.) *Machine translation* (Amsterdam, North-Holland, 1967), 77-135
Chomsky, N. (1957)
Syntactic structures. The Hague, Mouton, 1957
Dostert, L. (1955)
'The Georgetown-I.B.M. experiment', in: Locke, W.N. & Booth, A.D. (eds.) *Machine translation of languages* (Cambridge, Mass., M.I.T.Press, 1955), 124-135
Dreyfus, H.L. (1972)
What computers can't do: a critique of artificial reason. New York, Harper & Row, 1972
Eggers, H. (1981)
'Das Lemmatisierungssystem SALEM' *ALLC Bulletin* 9, 1981, 9-15
Garvin, P.L. (1972)
On machine translation: selected papers. The Hague, Mouton, 1972
Grishman, R. (1976)
A survey of syntactic analysis procedures for natural language' *American Journal of Computational Linguistics,* microfiche 47, 1976
Hutchins, W.J. (1978)
'Machine translation and machine-aided translation' *Journal of Documentation* 34, 1978, 119-159
Isabelle, P. et al. (1978)
'TAUM-Aviation: description d'un système de traduction automatisée des manuels d'entretien en aéronautique' Paper given at COLING 1978
Josselson, H.H. et al. (1972)
Fourteenth (Final) annual report on research in computer-aided translation Russian-English, Wayne State University, April 1972
Kay, M. (1973)
'Automatic translation of natural languages' *Daedalus* 102, 1973, 217-229
King, M. (1982)
'EUROTRA: an attempt to achieve multilingual MT', in: Lawson, V. (ed.) *Practical experience of machine translation.* (Amsterdam, North-Holland, 1982), 139-147
Kittredge, R. & Lehrberger, J., eds. (1982)
Sublanguage: studies of language in restricted semantic domains. Berlin, de Gruyter, 1982
Kulagina, O.S. et al. (1971)
Ob odnoi vozmozhnoi sisteme mashinnogo perevoda. Moskva, Inst. Russkogo Yazyka AN SSR, 1971
Lamb, S.M. (1966)
Outline of stratificational grammar. Washington, D.C., Georgetown Univ. Pr., 1966
Lehmann, W.P. & Stachowitz, R. (1972-75)
Development of German-English machine translation system. Final (annual) report(s). Austin, Univ.Texas, Linguistics Research Center, 1972(-75)
Luckhardt, H.D. (1982)
'SUSY: capabilities and range of application' *Multilingua* 1, 1982, 213-219

Maas, H.D. (1983)
'Zur Strategie der Analyse russischer Sätze', in: Herzog, R. (ed.) *Computer in der Übersetzungswissenschaft* (Frankfurt a.M., Lang, 1983), 63-73
Masterman, M. (1957)
'The thesaurus in syntax and semantics' *Mechanical Translation* 4, 1957, 35-43
Mel'chuk, I. A. & Zholkovskii, A.K. (1970)
'Towards a 'meaning-text' model of language' *Linguistics* 57, 1970, 10-47
Pendergraft, E.D. (1967)
'Translating languages', in: Borko, H. (ed.) *Automated language processing* (New York, Wiley, 1967), 291-323
Pigott, I.M. (1979)
'Theoretical options and practical limitations of using semantics to solve problems of natural language analysis and machine translation', in: MacCafferty, M. & Gray, K. (eds.) *The analysis of meaning; Informatics* 5 (London, Aslib, 1979), 239-268
Plath, W.J. (1967)
'Multiple-path analysis and automatic translation', in: Booth, A.D.(ed.) *Machine translation* (Amsterdam, North-Holland, 1967), 267-315
Sager, N. (1981)
Natural language information processing; a computer grammar of English and its applications. Reading, Mass., Addison-Wesley, 1981
Simmons, R.F. (1973)
'Semantic networks: their computation and use for understanding English sentences', in: Schank, R.C. & Colby, K.M. (eds.) *Computer models of thought and language* (San Francisco, Freeman, 1973), 63-113
TAUM (1973)
Taum 73: Projet de Traduction Automatique de 1'Université de Montréal. Rapport, Jan. 1973 (microfiche)
Thouin, B. (1982)
'The METEO system', in: Lawson, V. (ed.) *Practical experience of machine translation* (Amsterdam, North-Holland, 1982), 39-44
Toma, P. (1977a)
'SYSTRAN as a multilingual machine translation system', in: *Overcoming the language barrier* (München, Vlg. Dokumentation, 1977), 569-581
Toma, P. (1977b)
'SYSTRAN: ein maschinelles Übersetzungssystem der 3. Generation' *Sprache und Datenverarbeitung* 1, 1977, 38-46
Van Slype, G. & Pigott, I.M. (1979)
'Description du système de traduction automatique SYSTRAN de la Commission des Communautés Européennes' *Documentaliste* 16, 1979, 150-159
Vauquois, B. (1975)
La traduction automatique à Grenoble, Paris, Dunod, 1975
Wilks, Y. (1973)
'An artificial intelligence approach to machine translation', in: Schank, R.C. & Colby, K.M. (eds.) *Computer models of thought and language* (San Francisco, Freeman, 1973), 114-151
Wilks, Y. (1975a)
'An intelligent analyzer and understander of English' *Communications of the ACM* 18, 1975, 264-274

Wilks, Y. (1975b)
'Preference semantics', in: Keenan, E. (ed.) *Formal semantics of natural language* (Cambridge, Univ. Pr., 1975), 329-348
Woods, W.A. (1970)
'Transition network grammars -for natural language analysis' *Communications of the ACM* 13, 1970, 591-606
Woods, W.A. (1973)
'An experimental parsing system for transition network grammars', in: Rustin, R. (ed.) *Natural language processing* (New York, Algorithmics Pr., 1973), 111-154
Yngve, V.H. (1957)
'A framework for syntactic translation' *Mechanical Translation* 4, 1957, 59-65
Yngve, V.H. (1967)
'MT at M.I.T. 1965', in: Booth, A.D.(ed.) *Machine translation* (Amsterdam, North-Holland, 1967), 451-523