

[From: *UEA Papers in Linguistics* 9, January 1979, pp. 29-52]

Linguistic models in machine translation

John Hutchins

The Library, University of East Anglia, Norwich

One characteristic of modern linguistics has been the attention given to the formalisation of descriptions of linguistic systems. While rigour and precision have been a feature of the writings of linguists since the late nineteenth century (e.g. the Neogrammarians), it is primarily the work of Chomsky (1957, 1965) which has placed formal grammar at the centre of theoretical linguistics. The basic argument, now well known, is that the principal task of linguistic theory is to describe the 'competence' of the 'ideal speaker-hearer' in a 'perfectly explicit' grammar satisfying certain criteria of 'adequacy' (Chomsky 1965: 3-5). Formalisation assumes that language is (at least potentially) a well-defined system – a view which, of course, not all linguists share (e.g. Hockett 1968) – and it assumes that explanations can be found (perhaps non-linguistic ones in the main) for discrepancies between the actual 'performance' of individual speakers and the hypothesized 'competence' of ideal speakers. Although a fundamental requirement of a formal theory is that it should go beyond the mere description of what the 'competent' speaker knows and should provide explanations for why the grammar is constructed as it is, all agree that the first test any formal grammar must pass is that of 'descriptive adequacy', i.e. that it can account for observed linguistic facts. It is somewhat paradoxical that in contrast to the amount of effort devoted to theoretical discussions there has been so little objective evaluation of the formal grammars which have been proposed.

There are a number of reasons for this neglect by transformationalists (and indeed by theorists adopting other linguistic models). One is the argument or assumption that since all speakers of a language must 'know' the 'rules' of the language system in order to speak it and they know when they make mistakes, they can function as reliable informants. Consequently, it is argued, linguists can rely on their own intuitions about what is correct ('grammatical') and what is not to provide the necessary testing of their own formalisations. But it is an assumption which needs to be tested; the reliability of individual intuitions about language usage cannot be taken for granted, as Labov (1975) has demonstrated, least of all perhaps the intuitions of a linguist pleading for his own particular explanation or description.

A second reason for linguists' reluctance to subject their models to objective testing is probably their full awareness of the fragmentary state of the work done so far. They would argue that they are still uncertain what kind of grammar is appropriate for natural language and what the general characteristics of the formal model should be. A related reason is that many linguists would be uncertain what tests could objectively confirm or disconfirm a particular formal grammar. It is obvious, for example, that methods can readily be devised to test whether a given set of rules does or does not generate sentences which the compiler has judged to be 'grammatical', e.g. by realising the grammar in a computer program (cf. Friedman et al. 1971). But even if such a test confirmed the adequacy of the rules it would not demonstrate the validity of the model against actual usage.

In consequence, much of the theorising in linguistics about the form of grammars and about the formal treatment of particular linguistic phenomena (case

relations, semantic features, transformational constraints, pronominalisation, passivization, etc.) is carried out in a vacuum with no direct contact with real linguistic data. How can it be known whether a formal grammar is 'adequate' if it is not tested? One test of a grammar is to see whether it can be used (or adapted) in some model involving the processing (analysis and/or production) of actual text. Machine translation provides a suitable context for such a test.*

In general linguists have tended to ignore problems of translation; the theory of translation is one of the least developed areas of modern linguistics. The common attitude can probably be summarised as: "we cannot yet describe linguistic processes involving one language only, let alone attempt to describe what goes on in translation". Why then should machine translation be regarded as a suitable test-bed for linguistic theory? The principal reason is that whether a text produced by a machine translation (MT) system is or is not a reasonable translation of another text in another language can be evaluated by independent judges. "It provides a clear test of the rightness or wrongness of a proposed system... since the output in a second language can be assessed by people unfamiliar with the internal formalism and methods employed" (Wilks 1975a). The evaluation of translations has its problems, but in principle it can be objective, e.g. by observing whether the users of a manual produced by MT can understand and carry out instructions as well as users of versions of the manual produced by human translators (Sinaiko & Klare 1972), or by making back-translations of a MT text into the original language and looking at the differences – a test which can be done by someone knowing only the original language (Brislin 1976).

There are probably many reasons why linguists have generally been unwilling to be associated with machine translation – ignorance of the ways of the computer, more interest in theory than in practical work, etc. – but often it has been from a mistaken conception of the real aims of machine translation. The primary stimulus for MT research has always been the urgent needs of scientists, engineers, technologists, economists, administrators, etc. to cope with an ever increasing volume of material in foreign languages. In the 1950s and 1960s most demand was for access to Russian scientific literature and most early MT systems were designed for Russian-English translation. More recently the administrative and executive needs of the European Communities and the bicultural policy of the Canadian government have stretched existing translation services beyond their capacities to meet the heavy demand for technical and legal translations. Rarely are high quality translations required, normally all that is needed by administrators and scientists is to know the general content of texts. In these circumstances a MT system which can produce rough 'imperfect' translations quickly and relatively cheaply becomes a viable economic proposition. There is no question of attempting to produce high quality translations of literary texts; the objectives of MT research are severely practical and realistic.

This essay is concerned with the linguistic aspects of MT research. It attempts to describe briefly the formal models adopted in MT systems and to assess their adequacy for the purposes of translation processes. It does not deal with the wider issues of MT research and its relations to other areas of computational linguistics and artificial intelligence. For a fuller picture of current MT activity and related work see

* Obviously, no single kind of test can prove the general validity of a grammar, it can only provide supporting evidence. On the other hand, tests can demonstrate the inadequacy of grammars as general models (cf. Popper (1972) on the verification and falsification of theories).

Bruderer (1978) and Hutchins (1978), where bibliographical references for the systems to be described will be found.

It is unfortunate that the public image of MT has been formed by the disastrous and grossly expensive mistakes of the early work on MT. There is perhaps no other scientific enterprise in which so much money has been spent for so little return. By 1965 it has been estimated that U.S. government agencies had supported MT research at 17 institutions to the tune of almost 20 million dollars (Roberts & Zarechnak 1974). A sudden and abrupt end came with the report of the Automatic Language Processing Advisory Committee (ALPAC), set up by the National Science Foundation at the instigation of the U.S. sponsoring bodies, which concluded that MT was slower, less accurate and twice as expensive as human translation, that “there is no immediate or predictable prospect of useful machine translation” and furthermore that there was no shortage of technical and scientific translators in the United States (ALPAC, 1966). Although the report was widely condemned as narrow, biased and shortsighted, the damage had been done; henceforth MT was to be regarded, not least by linguists, as an expensive failure and anyone seriously advocating research in this field was to be looked upon as eccentric and misguided (if not worse).

The negative conclusions of ALPAC are not surprising when we examine the MT systems which were in operation or under development at the time; from the linguistic point of view they were crude and naive in the extreme, and not only in hindsight: many writers at the time criticised MT researchers for the lack of sound linguistic theory in their systems, indeed in some cases for ignoring linguistic research altogether. Many of the earliest MI systems adopted a crude ‘word-for-word’ approach to translation. Words of the text to be translated, the source language (SL) text as it is commonly called, were looked up in a bilingual dictionary; the equivalent words of the target language (TL) were selected; some simple rearrangements of word order were performed; and the results were printed out. A typical example was the Mark II system for Russian-English translation installed in 1964 at the Foreign Technology Division of the U.S. Air Force and in use until 1970 (Kay 1975). It was the unfavourable reports of Mark II which were largely responsible for ALPAC’s recommendations (cf. the various appendixes in ALPAC 1966).

The general strategy employed in nearly all MT systems until the late 1960s was the ‘direct translation’ approach: systems were designed in all details specifically for one pair of languages, i.e., in most cases, for Russian as SL and English as TL. The basic assumption was that the vocabulary and syntax of SL texts should be analysed no more than necessary for the resolution of ambiguities, the correct identification of appropriate TL expressions and the specification of TL word order. Syntactic analysis aimed at little more than the recognition of word classes (verbs, nouns, adjectives, etc.) to discriminate homonyms, e.g. *control* as verb or noun; and semantic analysis (if it was included) was restricted to the use of features such as ‘male’, ‘concrete’, ‘liquid’, for resolving collocational ambiguity, e.g. in ‘The crook escaped’ *escape* specifies an ‘animate’ subject and thus excludes the inanimate sense of *crook* (‘shepherd’s staff’).

The Georgetown University system was typical of the ‘direct’ approach, and it proved to be the most successful of them all. For many years the MT research group at Georgetown under Léon Dostert was the largest in the United States (Kay 1975, Dostert 1963). In 1964 the group delivered an operational Russian-English system to the Atomic Energy Commission at Oak Ridge National Laboratory and to the EURATOM centre in Ispra, Italy; in both places the system was in regular use until very recently. The Georgetown system illustrates well the complexities and the

ultimately insuperable problems of the 'direct' approach. Despite a monolithic grammar of "monstrous size and complexity" its syntactic analysis was very rudimentary, devoted to nothing more than resolving problems in the assignment of word-classes. The methods were *ad hoc*, there was no notion of grammatical rule or of syntactic structure. "Such information about the structure of Russian and English as the program used was built into the very fabric of the program so that each attempt to modify or enhance the capabilities of the system was more difficult and more treacherous than the last" (Kay 1975). Indeed the systems at both Oak Ridge and Ispra remained virtually unchanged since their installation. Although undoubtedly the translations produced were poor, the users seem to have been well pleased (Dostert 1973). The results were not unreadable and, with some knowledge of the subject matter, scientists were able to extract the information they needed. Quite clearly, they would much rather have a low quality MT product than have no translation at all.

From the Georgetown approach has emerged the only MT system at present in full operation. This is SYSTRAN (Toma 1977), a 'direct' Russian-English translation system, which has also been adapted for English-French translation (hence the interest of the Commission of the European Communities). From the linguistic standpoint, SYSTRAN represents little advance on its Georgetown 'ancestor'. The main improvement lies in the 'modularity' of its programming, allowing modifications of any part of the translation processes to be undertaken without fear of impairing overall efficiency. Furthermore, the linguistic and computational facets are kept separate, thus avoiding the irresolvable complexities encountered in the Georgetown systems.

In SYSTRAN there are four basic stages in the translation process: Input, Dictionary lookup, Syntactic analysis, Translation. After the preparatory Input stage, each word of Russian text is checked against two dictionaries (first the High Frequency dictionary, then a Master Stem dictionary) for information on grammatical (and some semantic) properties and for possible English equivalents. Syntactic analysis involves four 'passes': first to resolve homographs, then to establish basic phrase groups (verb plus object, preposition plus object, etc.), then to extend phrase structures and identify specified objects and complements, and lastly to determine the types of clauses (e.g. subordinate), their ranges and their constituents (subjects and predicates). It provides at most a rudimentary immediate constituency description, plus the identification of certain basic grammatical relations (subject, predicate, object). The final stage, Translation, consists of many subroutines using information from the dictionaries and from the syntactic analysis for the selection and arrangement of the English output. There is no consistent methodology; any information leading to acceptable English text is employed whatever its source. For example, the routine to insert definite and indefinite articles combines syntactic information (e.g. whether the Russian noun is qualified by a following genitive noun, prepositional phrase or relative clause), semantic information (e.g. whether the Russian is an ordinal number) and information on English equivalents (e.g. English 'mass' nouns usually require definite articles). In some cases English syntactic form is determined by codes in Russian lexical items, e.g. ESLI includes a code to change a Russian infinitive construction ('if to examine...') to an English finite form ('if we examine...'); in other cases English syntactic form results from a manipulation of the output, e.g. 'noun + *of* + verbal noun + *of* + noun' (*result of treatment of burns*) becomes 'noun + *of* + gerundive + noun' (*result of treating burns*). Elsewhere an *ad hoc* system of 'semantic classification' is employed, e.g. the translation of Russian prepositions according to the 'semantic class' of adjacent verbs or nouns; but these classifications vary from one subroutine to another and they have little to do with the semantics of

Russian, they are usually merely labels designed to overcome particular difficulties with English output.

The linguistic model underlying SYSTRAN is clearly not based on a particular theory of grammar or of translation; in this respect it is much less sophisticated than later developed MT systems. Yet despite this linguistic 'crudity' it has to be acknowledged that SYSTRAN does actually work and it is producing quite acceptable translations, as the following example illustrates (Toma et al. 1974):

The question concerning the semantic interpretation of the models of a sentence is one of the most complex questions of the modelling theory of sentences. The multiple attempts at the semantic substantiation of the models of sentences, which took place of the purely structural classification of the models of the sentence of the descriptivists, as well as the serious criticism of these attempts and the calls to deny the semantic interpretation of the models of a sentence are known.

The practical success of SYSTRAN is one argument for linguists becoming more familiar with its procedures, particularly since (as we shall see) MT systems based on more sophisticated linguistic models have not so far proved any more successful and have in fact in a number of cases proved to be failures. The other argument for linguists paying more attention to SYSTRAN is that in the course of its development much valuable information has been accumulated on the syntax and vocabulary of Russian and English (and more recently of French also) from a large number of texts of many thousands of words. This information appears to remain unexploited by the linguistic community at large.

Since the mid 1960s and the ALPAC report research on MT has been both more circumspect in its claims and ambitions and more attentive to developments in theoretical linguistics. Whereas in the 'first generation' of MT systems the approach was essentially that of the engineer seeing problems of MT as practical technical difficulties to be overcome by trial and error (the 'brute force' approach, as Garvin (1972) characterised it), in the last decade MT systems have been based on clearly articulated linguistic models. In this 'second generation' of MT research, the 'direct' approach has generally been abandoned in favour of 'interlingual' or 'transfer' approaches. Translation is indirect via an intermediary language (interlingua) or via a transfer component operating upon 'deep syntactic' or semantic representations. Whereas in 'direct' systems the analysis of SL texts is determined by the requirements of TL text production, in 'interlingual' and 'transfer' systems the analysis of SL texts is quite independent of the TL. The systems are not therefore designed for translation only between two specific languages but can in principle be adapted for translation between other pairs of languages by the addition of new programs of SL analysis and TL synthesis.

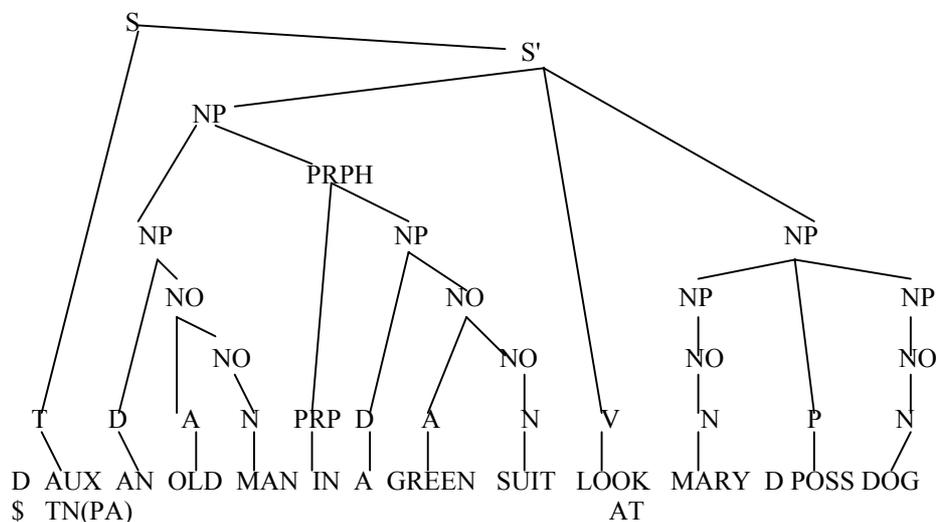
The attractiveness of an interlingual approach to MT was mentioned as early as 1949, by Warren Weaver in the memorandum sent to some two hundred acquaintances which effectively launched MT as a scientific enterprise (Weaver 1955). But it was not until the 1960s when theoretical linguistics had turned to problems of language universals that MT researchers had any clear ideas of how interlinguas could be constructed.

Transformational-generative grammar provided one obvious model. The best example of MT research based on this approach was the work at the University of Texas in a team under Lehmann and Stachowitz (1972-75). The aim was to develop a MT system for German-English translation (called METALS) which could also be adapted to other pairs of languages. At the time when the group began in the early

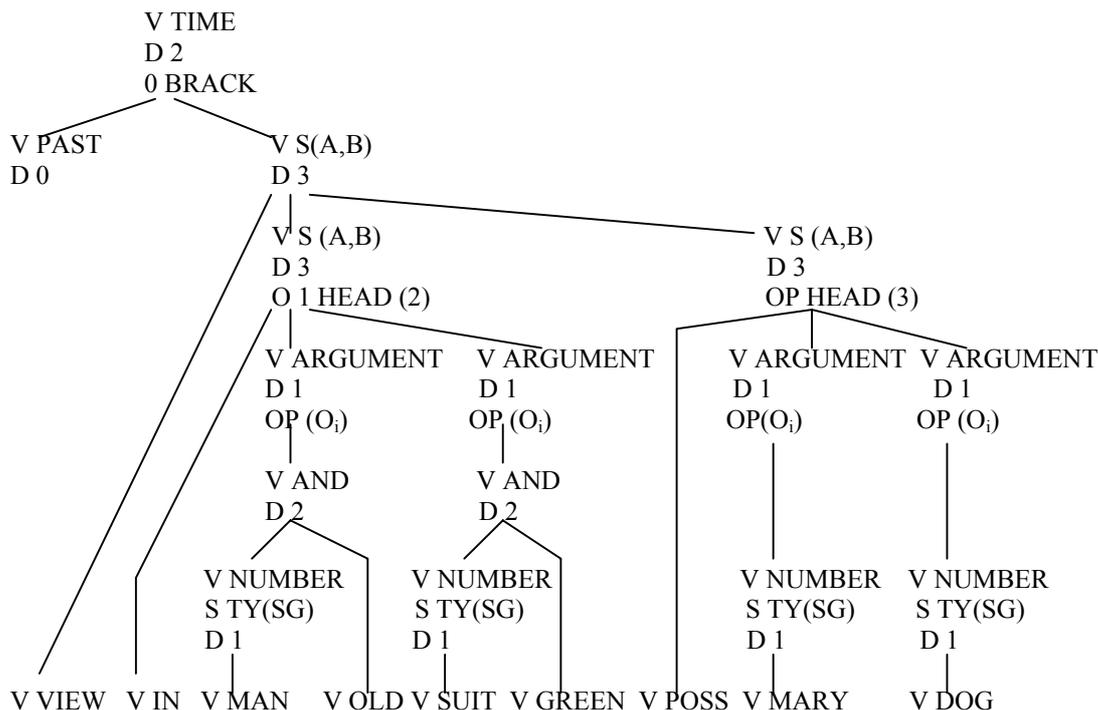
1960s it was argued that while languages differ in ‘surface structures’ they all share the same ‘deep structures’ and that, since transformational rules do not affect the meanings of sentences, deep structures may be regarded as forms of ‘universal’ semantic representations.

This conception of deep structure has been virtually abandoned in later developments of TG theory, as is well known, but it is interesting to find that MT researchers also encountered difficulties with the *Aspects* model (Chomsky 1965), although for different reasons. The METALS team discovered soon that the kind of transformational rules formulated in TG theory could not be implemented without considerable complexities in programming. As other groups also learned, 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 eliminate information from deep structures and there is no way this information can be reconstructed with certainty (cf. Grishman (1976) for discussion of parsers). The METALS team adopted therefore a conception of transformation closer to that of Harris (1957).

The METALS interlingua was not a genuine interlingua. It was restricted to syntactic structures, into which and from which German and English sentence forms could be analysed and synthesized. There was no attempt to decompose lexical items, e.g. into semantic primitives; conversion of vocabulary items from German to English was made through a normal bilingual dictionary. Hence it could not even be truly universal as a syntactic interlingua; it could not handle such semantic equivalences as *He ignored her* and *He took no notice of her* since they would have different deep structures. Analysis was performed by three ‘grammars’ working in sequence. After morphological analysis and dictionary lookup, the ‘surface sentence’ was converted by a ‘surface grammar’ into one or more tentative ‘standard strings’. In this process certain elements discontinuous in the surface form (e.g. verbs such as *look ... up*) would be brought together. In the second stage, the tentative ‘standard strings’ were tested by a ‘standard grammar’ for syntactic well-formedness and each string accepted by the ‘standard grammar’ was then provided with one or more phrase-structure representations, called ‘standard trees’. The result of such an analysis for the sentence *An old man in a green suit looked at Mary’s dog* is illustrated in the standard tree below (from Lehmann and Stachowitz 1972):



The third stage, 'normalization', filtered out semantically ill-formed standard trees by testing the semantic compatibility of syntactically related lexical items (referring to information provided by the dictionary), i.e. much in the way proposed for semantic interpretation in the Standard Theory of transformational grammar. Each standard tree accepted was then converted into a 'normal form' (or several 'normal forms' if it was ambiguous), a 'deep structure' representation in which the relationships between items were expressed in terms of 'predicates' and 'arguments', i.e. in this respect rather like the Generative Semantics conception of deep structure. For the standard tree above would be derived the following normal form:

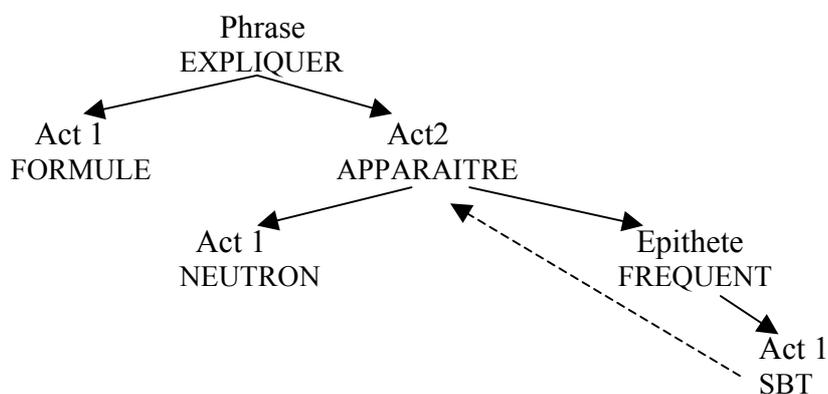


Synthesis of target language (English) sentences proceeded first by the substitution of TL lexical elements in the 'normal form', then the production of 'standard strings' and finally by the conversion of strings into 'surface sentences'.

Like many other MT systems, METALS suffered from an inadequate method of syntactic analysis. The phrase structures produced by the 'standard grammar' were derived entirely on the basis of grammatical information assigned to elements within single sentences. The result was inevitably that multiple analyses were provided for most sentences. For example, since a prepositional phrase may be governed by either a verb or a preceding noun phrase, a sequence such as $V + NP_1 + P + NP_2 + P + NP_3$ may receive parsings which relate NP_2 and V , or NP_2 and NP_1 , or NP_3 and V , or NP_3 and NP_1 , or NP_3 and NP_2 , or any possible combination of these analyses. Without semantic information it is not possible to decide whether in the sentence *They threw the boy in the river* the prepositional phrase *in the river* modifies *threw* or *the boy*. The absence of intersentential or discourse analyses also allowed multiple 'normal forms' to be produced for single 'standard trees'. Furthermore, since a single normal form could obviously be the source of many different (but semantically equivalent) surface forms the problems of synthesis were also multiplied. As a MT system it was clearly unsatisfactory. However, from the beginning the system had been conceived as a general-purpose system designed also for other automated language processes

and in later years (the project ended in 1975) most emphasis was placed on research connected with automatic indexing and abstracting.

The Centre d'Études pour la Traduction Automatique (CETA) at the University of Grenoble began research on a Russian-French MT system in 1961, at about the same time as the Texas team. It too adopted the interlingual approach, but the underlying linguistic theory was different. As in METALS the first stages of analysis in CETA were Dictionary lookup and Morphological analysis, followed by a phrase structure analysis (at which stage discontinuous surface forms were brought together). The next stage, however, converted the resulting 'surface syntactic' structures into dependency-tree representations. First the phrase structures were augmented by dependency relations, so that, for example, in a VP the V was marked as 'governor' and the NP as 'dependent'. Then the lexical items were classed as either predicatives or non-predicatives, where predicatives included adjectives and adverbs as well as verbs and where non-predicatives were nouns and articles. Next the structures were analysed in terms of predicatives and their arguments (non-predicatives or other predicatives), resulting, after the removal of word-classes (N, V, Adj, etc.), in a tree such as:



(where Act = 'actant' (cf. Tesnière 1959) and where SBT stands for the argument dependent on FREQUENT, i.e. APPARAÎTRE). At the same time, as in METALS, semantically anomalous analyses were 'filtered out' by checking the compatibilities of constituent elements from information supplied by the dictionary (e.g. data on 'selectional restrictions').

Such a tree was the source for TL synthesis. First, lexical units of the SL were replaced by equivalent TL units; then these units were examined for their potential word-classes and dependency relations; a predicative was located and its arguments checked as possible NP dependents (if one argument was itself a predicative, e.g. APPARAÎTRE in the tree above, then the possibility of a clause structure was also investigated, i.e. ...*que...apparaître* as well as NP *apparition*); then argument nodes (Act 1, Act 2, etc.) were replaced by appropriate categories (V, NP, Adj, etc.), and the elements were reordered to conform to TL syntax; finally morphological synthesis completed the process by producing the correct surface forms (including the editing of variants, e.g. *le* → *l'* before *a, e, i, o, u.*)

Interlingual (or 'pivot language') representations such as the one above show that the CETA model has some affinity to the dependency grammar of Tesnière (1959). But it has been influenced more directly by the 'meaning-text' model of the Russian linguist Mel'chuk (Mel'chuk & Zholkovskii 1970), as the principal designer of CETA, Vauquois (1975), has acknowledged. Mel'chuk's model is stratificational

in conception, recognising four basic levels of linguistic representations and a system of 'grammars' for converting representations from one level to another. Like Lamb's analogous but nevertheless quite distinct and independent stratificational model (e.g. Lamb 1966), Mel'chuk's original conception had 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 theoretical speculations. The 'strata' of Mel'chuk's model are: phonemic representation, morphological representation, surface syntactic representation (including grammatical relations such as 'complement-of', 'subject-of', 'auxiliary', 'determinant', indication of anaphoric relations, structure of nominal groups, and theme-rheme relations), deep syntactic representation (tree structures composed of valency relations among root lexical elements (sememes) and incorporating anaphora and informational (given-new), focus and theme-rheme indicators), semantic representation (network structures of abstract semantic relations among semantic primitives which correspond to a number of possible 'deep syntactic representations'). In a MT system it was recognised that analysis of SL text need not go as far as a full semantic representation; instead, transition from a SL deep syntactic representation to a TL deep syntactic representation was achieved by a series of 'paraphrasing' operations.

Insofar as CETA's SL-TL conversion was at the level of 'deep syntactic representation' which has some correspondence to this level in Mel'chuk's model it may be said to be a MT implementation of 'stratificational' grammar. However, CETA lacked the detailed paraphrasing operations present in Mel'chuk's model, which involve not only lexical relations (synonymy, nominalisation, adverb formation, causative/inceptive/terminating/factitive (etc.) relations) but also syntagmatic structural equivalences (cf. Zholkovskii & Mel'chuk 1970). It is true that CETA does conflate certain semantically equivalent syntactic structures (e.g. in the above figure the subtree dominated by APPARAITRE as a noun phrase and as a subordinate clause), but like METALS it cannot deal with equivalences involving different lexical formations; and it is precisely such phraseological equivalences that the paraphrase operations of Mel'chuk's model are designed for. More importantly perhaps, CETA did not retain information about theme-rheme, choice of subject noun, use of passive, subordination of clauses, etc.; such information about the 'surface' forms of SL text could help considerably in the selection of appropriate TL forms. Above all, the system as a whole was too rigid: if morphological analysis failed because the dictionary had no entry for a particular word or did not record all homographic variants, then this affected all subsequent processes; if syntactic analysis failed to parse any part (however small) of a sentence, it was rejected. In addition, like the METALS parser, too many analyses were attempted which came to nothing and too many analyses were produced which had to be 'filtered out' later. What was needed was a parser which did not use its full armoury of analytical techniques for every simple phrase structure but reserved the more complex parts for only complicated sentence structures.

Experience with linguistically ambitious MT systems like METALS and CETA has led to the adoption of more modest 'transfer' approaches. It seems at present to have been conceded in MT research that the 'pure' interlingual approach is not feasible, at least not until linguistic theory has advanced much further in the study of language universals. It is true that neither the METALS nor the CETA systems were in fact full interlingual systems, particularly in their semantic aspects, nevertheless the ultimate aim was to develop 'deep structure' representations embodying what was common to two languages and hence to make the first steps towards 'universal'

representations. In the 'transfer' approach, there is no intention to provide semi- or quasi-universal 'deep structure' representations. The goal of analysis programs is to produce representations of sufficient abstractness (or 'depth') to enable the compilation of reasonably simple 'transfer components'. Whereas in the interlingual approach translation is a two stage process (Analysis of SL text into Interlingua and Synthesis of TL text from Interlingua) in the transfer approach it is a three stage process: Analysis of SL text into SL 'deep' representation, Transfer from SL 'deep' representation to TL 'deep' representation and Synthesis of TL text from TL 'deep' representation. For any particular language the programs of SL analysis and TL synthesis are held constant whatever the other language involved; only the Transfer programs are specific to particular language pairs. Obviously, the more abstract the 'deep' representations can be and the simpler the Transfer programs, the greater advantage such MT systems will have over the 'direct' systems such as SYSTRAN. On the other hand, the more the designers of 'transfer' systems can avoid or circumvent the problems of language universals and the intricacies of detailed semantic and pragmatic analysis the more certain they will be that the resulting system will work in practice. In consequence, analysis rarely goes further than the familiar territory of syntax, and there is still little use of semantic analysis.

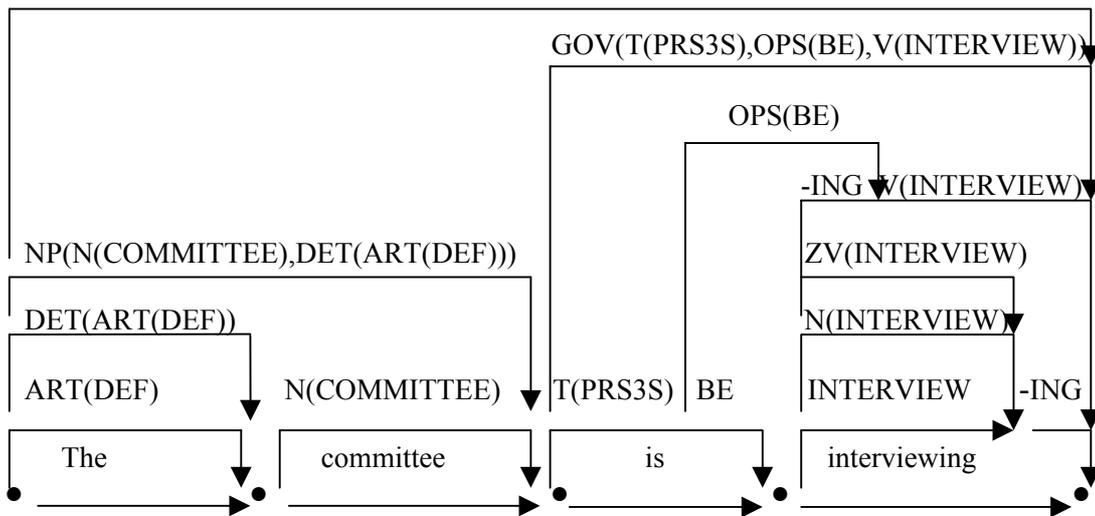
The best example of a 'transfer' MT system is the TAUM project at Montreal (TAUM 1973). It is also the 'second generation' system which is nearest to full operational implementation, as Projet Aviation, an English-French MT system for the Canadian Air Force (TAUM-Aviation 1977). In TAUM translation proceeds in five stages: Morphological analysis of English, Syntactic analysis of English, Transfer, Syntactic generation of French, Morphological generation of French. Each stage consists of a grammar of 'Q-systems'; Q-systems are computer programs designed to manipulate tree structures and strings of trees irrespective of the labels attached to the nodes of trees. A tree may be a phrase-structure representation, e.g. PH(SN(IL), SV(V(MANGE), SN(LA,CHOUCROUTE))), or it may represent a list (items separated by commas), e.g. L(A,B,C,D), where each item may itself be a tree, or it may represent a categorisation, e.g. PREP(TO) or a single node, e.g. TODAY. A string of trees is defined as a sequence of trees separated by plus signs, e.g. SN(PAUL) + V(VIENDRA) + DEMAINE + CHEZ + PRON(MOI). A Q-system rule converts strings (of one or more trees) into new strings, it may apply to the whole or to only a part of a string, and it may include variables for labels, lists or trees. For example, in the rule $PREP(A^*) + SN(X^*) \rightarrow OBJIND(P(A^*),SN(X^*))$ the A^* is a variable for a label (TO, FROM, ...) and the X^* is a variable for a list (of nouns). Clearly, the Q-system formalism is very powerful, capable it would seem of handling morphological and syntactic representations within any formal model.

Morphological analysis involves the assignment of category labels (e.g. WITHIN \rightarrow P(WITHIN)), segmentation of prefixes (e.g. UNDERSTOOD \rightarrow UNDER + STOOD), regularisation of irregular forms (e.g. STOOD \rightarrow SW(STAND) + ED(PST)), identification of suffixes (e.g. TRIED \rightarrow TRI + ED, PUTTING \rightarrow PUTT + ING), construction of base forms (TRI \rightarrow TRY, PUTT \rightarrow PUT). Dictionary lookup includes the assignment of category labels (ADJ, N, ...) and 'features' (e.g. ANI, CONC, ABST for nouns, features of admissible arguments (subject nouns, objects, etc.) for verbs).

Syntactic analysis is in two stages. The first includes the recognition of noun phrases and complex verb forms and the rearrangement of constituents as needed, e.g. $DET(V^*) + N(X^*) \rightarrow NP(N(X^*),DET(V^*))$. The second establishes the 'canonical form' of sentences. It incorporates both phrase structure rules and transformational

rules: input strings of trees are formed into single complex trees and reordered (or deformed) as ‘deep structure’-type representations. Thus, verbs are put before their argument noun phrases, passive constructions are made active, extraposed *it* forms are transformed (e.g. It be ADJ that S → S be ADJ) and relative pronouns are replaced by REL and the head noun copied into its argument position in the subordinate clause. An example of a TAUM analysis is the following. Each arrow line represents a step in the analysis (i.e. the application of a replacement rule) working upwards from the ‘surface form’ at the bottom to the final form at the top.

IX(GOV(T(PRS3S),OPS(BE),V(INTERVIEW)),NP(N(COMMITTEE),DET(ART(DEF))))



Transfer involves the translation of English ‘words’ with their category labels into French equivalents via a bilingual dictionary and the modification of certain parts of trees to simplify generation. In Syntactic generation successive Q-systems break down the complex tree output from Transfer into strings of trees, e.g. the noun phrase SN (N (GENS), DET (LES), GP (P (DE), SN (N (VILLAGE), DET (LE))) becomes DET(LES) + N(GENS) + P(DE) + DET(LE) + N(VILLAGE). Finally, morphological generation converts trees and strings into single ‘surface’ forms (e.g. DET(LES) → *les*, P(DE) + DET(LE) → *du*).

The TAUM system illustrates well characteristic features of ‘transfer’ MT systems: the clear separation of the different stages of analysis and synthesis, the separation of linguistic data from the processing algorithms (in this case, Q-systems), and the use of separate dictionaries for SL analysis, transfer and TL synthesis. The separation of stages is now generally regarded as essential in all MT systems if the programming is to be kept under control: the ‘modularity’ of SYSTRAN is perhaps its principal improvement over the Georgetown system. Likewise, the necessity of keeping apart the linguistic information (e.g. the rules of formal grammar) and the programming algorithms is now universally accepted. Lastly, the computational advantages of separate smaller, less complex dictionaries over the monolithic bilingual dictionaries of earlier ‘direct’ systems are also accepted by most current MT researchers.

As a linguistic model TAUM is obviously less complete than either CETA or METALS; there is hardly any semantic analysis for the resolution of residual syntactic ambiguity and even the syntactic analyses are not as ‘deep’ (in TG terms) as those in CETA and METALS. In these respects TAUM illustrates a strong feeling among many

MT researchers that the approach to linguistic modelling adopted or assumed by the 'pure' linguistic theorists is not appropriate. For the practical objectives of producing quick translations of technical documents it may be better to take a more pragmatic stance: to use the computer to do only what it can do well, accessing large dictionaries, making morphological analyses and producing simple 'rough' parsings, and to use human skills for the more complex problems of semantic analysis, resolving ambiguities and selecting the appropriate expression from a choice of possible translations. In recent years there has thus been a number of 'interactive' MT systems under development. One example is the MIND system (Kay 1973), a general purpose language data processing system designed to carry out a great variety of tasks including grammar testing and question answering as well as translation. Its components are morphological and syntactic analysers, semantic file processor, transformational component, morphological synthesizer and interactive disambiguator. As a MT system, MIND takes the form of a 'transfer' system with human collaboration. After a sentence has been automatically analysed as a 'deep structure' representation by the morphological and syntactic components, it is presented to the human consultant for the resolution of ambiguities, e.g. problems concerning prepositional phrases or homonyms. The interactive disambiguator decides what the problems are and what questions need to be answered to resolve them. Given a sentence such as *They filled the tank with gas* it might ask:

	DOES THE WORD 'TANK' REFER TO	DOES 'GAS' REFER TO
	1. A MILITARY VEHICLE?	1. GASOLINE?
	2. A VESSEL FOR FLUIDS?	2. VAPOR?
or:	DOES 'THEY' REFER TO	
	1. SOLDIERS?	
	2. TANKS?	
	3. SHELLS? (or any other recently used noun)	

In the case of a sentence such as *He saw the girl with the telescope* it might ask:

DOES THIS MEAN
1. 'SAW WITH THE TELESCOPE'?
2. 'GIRL WITH THE TELESCOPE'?

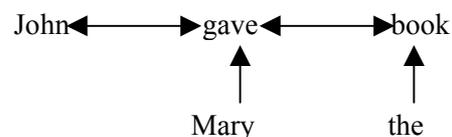
It is the resolution of such difficulties, simple enough for the human translator, which cause such considerable problems for fully automated systems. Furthermore these are problems primarily of analysis; once they are resolved the synthesis of TL texts is relatively easy. Consequently interactive systems are most attractive where there is a need for simultaneous translation of a single SL text into a number of languages; the expensive involvement of a skilled human analyst is then employed to the greatest advantage.

Interactive systems are one answer to the inadequacies of current models. Another, more radical, solution is to abandon syntax-based models and to adopt semantics-based methods of analysis. All the MT systems described so far are essentially syntax-based: however much semantic information is included in intermediary representations, syntactic analysis is the central component; semantic analysis operates only after syntactic structures have been determined. As the CETA researchers discovered, failure in the syntactic components cannot be overcome by the semantics component, however sophisticated. The systems are also syntax-based in another sense: their analytical procedures are restricted to sentences. Few systems are able to deal at all with cross-sentence pronominalisation (e.g. *they* in the MIND example above) and semantic links between sentences – those features which make a sequence of sentences into a cohesive whole (Halliday and Hasan 1976) – have been neglected.

The importance of Wilks' work on a prototype MT system lies precisely in the exploration of a semantics-based approach to analysis and in the incorporation of semantic and pragmatic text analyses. His framework is the research on language understanding by workers in Artificial Intelligence and the basic approach is 'interlingual'; Wilks describes a system for English-French translation (Wilks 1973, 1975a). The first stage is a fragmentation routine which partitions text at punctuation marks and specified keywords (prepositions, conjunctions, etc.) rendering for example *I advised him to go* as '(I advised him)(to go)'. Next each fragment is tested against an inventory of 'templates', semantic frames expressing the 'gists' of (parts of) sentences in the form of 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 interlingual elements or 'semantic primitives' which would be found as the principal ('head') semantic categories in the semantic formulas representing *John*, *own* and *car* respectively.

Semantic formulas are constructed from a limited number of 'elements' and left and right brackets, e.g. *drink* has the formula: ((*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 entries goes no further than necessary for the purpose; in this context there is no need to distinguish *mouth* from other apertures. The notion of preference is a central features of Wilks' method of analysis: SUBJ displays the preferred agents of actions and OBJE the preferred objects or patients, they do not stipulate obligatory features of agents and patients (as in syntax-based systems incorporating TG-type 'selectional restrictions') and thus they permit 'abnormal' usages (e.g. cars drinking petrol) while still expecting the 'normal'. In this way Wilks' "preference semantics" can cope with many types of metaphorical expressions (Wilks 1975b).

In the next stage, elements of fragments not so far included in templates are examined for their relationships to those already identified; thus, adverbs are linked to 'actions', adjectives to 'agents' or 'patients' and so forth. The result is a dependency network, e.g.



Then the program searches for dependencies between fragments, e.g. a temporal phrase (*during the war*) might be tied to the 'action' element of an earlier fragment or to the 'action' element of the following fragment, by a 'location' link. Such ties are made not only within sentences but also across sentence boundaries, since the basic unit of analysis is not the sentence but the phrase (fragment). Some ties involving pronominal reference make use of 'common sense inferences'. For example, in *The soldiers fired at the women and we saw several of them fall* the linking of *them* to *women* rather than to *soldiers* is made on the basis of a 'common sense rule' stating that if an animate object is hit it is likely to fall.

The distinctive features of Wilks' analytical method are thus the use exclusively of semantic features in the 'parsing' of phrases, the use of preference

semantics and common sense inference rules, and the analysis of discourse relationships. At no stage is there any reference to syntactic structures or indeed to the boundaries of sentences. Grammatical categories such as noun and verb have no role, not even in the resolution of homographs: to identify the verbal sense of *father* in a sentence such as *Small men sometimes father big sons* the program needs only to find that the semantic formula with CAUSE as its 'head' is the only one which will fit the other 'heads' in an acceptable template. In Wilks' system then semantic representations are reached without recourse to previous syntactic analysis.

There are undoubtedly reservations about such an approach among MT researchers; it is not known how feasible it would prove to be in a full-scale operation; the complexity and amount of information needed in semantic formulas and the difficulties of formulating 'common sense rules' have yet to be investigated. Nevertheless, it is now widely accepted in MT circles that future systems must incorporate components along the lines of Wilks' semantic parser, preference semantics and inferential semantics. In what form and to what extent they will actually be needed in practical MT systems is still very much an open question.

It is therefore encouraging to see the development in recent years of a MT framework which has the requisite flexibility to test alternative approaches to linguistic analysis. This is the GETA system (Boitet 1977) which is being developed at Grenoble University as the successor to their CETA system already described. Experience with the 'interlingual' CETA had revealed disadvantages in reducing texts to semantic representations and destroying in the process a good deal of 'surface' information useful for TL synthesis. There is no point, for example, in converting a SL passive form into an active representation if it has only to be reconverted into a similar TL passive form. The GETA system is highly flexible both in its programming and in its linguistic aspects, and it is designed to promote cooperative activities with other MT research groups.

GETA is basically a 'transfer' system with morphological and syntactic analysis, transfer, and syntactic and morphological synthesis, but the analysis goes much further than in TAUM. The results of the analysis programs are dependency-tree type 'deep structure' representations rather like those in CETA, i.e. aiming for language-independent semantic 'pivot language' representations. However, it is no longer (as in CETA) the objective to establish 'universal' pivot languages, rather each SL has its own 'pivot'. The Transfer program has two stages: the conversion of SL 'lexical' elements into equivalent TL 'lexical' elements (involving reformation of tree structures as necessary), and the conversion or transformation of SL 'pivot' structures indicating dependency relations into equivalent and appropriate TL 'pivot' structures. In a sense the GETA system is a flexible conjunction of the 'interlingual' and 'transfer' approaches.

The principle source of GETA's flexibility, however, lies probably in its algorithmic features. The major premise of the GETA team has been that the algorithms employed at any particular stage should be no more complex and no more powerful than necessary for handling the linguistic data in question. On this argument it rejects the use of such powerful algorithms as the Q-systems (of TAUM) and the 'augmented transition network' parsers (developed by Woods (1970) and others) for the simple manipulation of strings in, for example, morphological analysis and synthesis. For syntactic and semantic analysis the team has developed an algorithm for the transformation of one abstract tree or subtree into another, where the linguist decides what transformations are to be used in particular instances and what conditions are to be attached to their use. The linguist 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 adopt one kind of

linguistic model in one set of subgrammars and another model in another set of subgrammars, specifying the conditions for switching from one strategy to another. The system provides the linguist with a vast choice of approaches and assures him that, whatever the strategy or 'grammar' used, there will always be a result at the end of a finite application of rules. Unlike the earlier systems (including its predecessor), GETA does not test for the *acceptability* of structures (i.e. the subgrammars do not filter out ill-formed structures) but tests for the applicability of rules of transformation. The subgrammars work on sub-tree specifications, if a rule does not apply the tree remains unchanged; even if no rule of a subgrammar can be applied there will always be a tree as output on which other subgrammars may operate.

The flexibility of GETA offers the linguist the prospect of genuine tests of alternative linguistic models on actual real-life linguistic data (texts to be translated). There is no reason to doubt that GETA could not easily incorporate a 'semantic parser' or expand (or modify) its semantic information to include 'preference' and 'inference' semantics on the lines indicated by Wilks. There is equally no reason to think that GETA could not include 'grammars' based on approaches other than the transformational-generative and dependency models, e.g. building upon Winograd's (1972) experience with systemic grammar. Above all, perhaps, the GETA framework offers the prospect of fruitful cooperative efforts on multilingual MT research. A number of MT teams from France, Germany and Brazil have formed the LEIBNIZ group, adopting a common conceptualisation of MT system design and collaborating in the development of analysis and transfer programs (Boitet 1977).

The advances in computational and linguistic sophistication in recent years justify perhaps the mood of quiet optimism which is evident from the publications of current MT research groups. Nevertheless, it has to be admitted that these systems have yet to prove themselves. It is still a fact, embarrassing as it may be perhaps, that the only currently operational MT system is SYSTRAN, based on a linguistic design which owes almost nothing to the linguistic theory of the past twenty years. The ad hoc pragmatism of SYSTRAN has produced reasonable Russian-English translations of scientific and technical texts, while the more linguistically advanced 'second generation' systems have either been abandoned or have yet to move beyond the laboratory stage. It could well be that MT research has been 'led astray' by the kind of linguistic models which have been proposed in recent years. The inadequacies of syntax-based grammars, the lack of fully articulated semantic theories, and the need for adequate discourse and intersentential mechanisms have become more and more apparent. What is the reason for this 'failure' of linguistic theory to provide or suggest models appropriate to MT and other language processing systems? One may be the crucial assumption which distinguishes between 'competence' and 'performance'. Linguistic theory has concentrated on the formal definition of language systems and has neglected the investigation of language behaviour in social contexts; it 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 feasible has itself encouraged the dissociation of theory and practical reality which has led to the adoption of unrealisable models. If the present emphasis in MT research (and in artificial intelligence) on flexible process-oriented models with adequate semantic, discourse, inferential and knowledge-structure components produces successful systems – systems which pass the rigorous tests of practical translation – then the resulting 'model' (however 'impure' from a theoretical standpoint) deserves to be studied seriously by linguistic theorists. For too long too many linguists have devoted

themselves to the construction of 'ideal' systems; more attention should be paid to the practical needs of those tackling problems of 'raw' linguistic facts.

References

- ALPAC (1966): *Languages and machines: computers in translation and linguistics*. A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council. Washington, D.C., 1966
- Boitet, C. (1977): 'Où en est le GETA début 1977?' *T.A. Informations* 18, 1977, 3-20
- Brislin, R. W. (1976): 'Introduction' In: Brislin, R. W. (ed.) *Translation: applications and research* (New York, Garner, 1976), 1-43
- Bruderer, H. E. (1978): *Handbuch der maschinellen und maschinenunterstützten Sprachübersetzung*. München, Vlg. Dokumentation, 1978 (Forthcoming in English: *Handbook of machine translation and machine-aided translation*. Amsterdam, North-Holland Publ. Co., 1979)
- Chomsky, N. (1957): *Syntactic structures*. The Hague, Mouton, 1957
- Chomsky, N. (1965): *Aspects of the theory of syntax*. Cambridge, Mass., M.I.T. Pr., 1965
- Dostert, L. E. (1963): 'Machine translation and automatic language data processing' In: Howerton, P. W. (ed.) *Vistas in information handling* (Washington, D.C., Spartan Books, 1963), 92-110
- Dostert, B. H. (1973): Users' evaluation of machine translation, Georgetown MT system, 1963-1973. Final report. Texas A & M University, 1973
- Friedman, J. et al. (1971): *A computer model of transformational grammar*. New York, American Elsevier, 1971
- 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
- Halliday, M. A. K. and Hasan, R. (1976): *Cohesion in English*. London, Longman, 1976
- Harris, Z. S. (1957): 'Co-occurrence and transformation in linguistic structure' *Language* 33, 1957, 283-340
- Hockett, C. F. (1968): *The state of the art*. The Hague, Mouton, 1968
- Hutchins, W. J. (1978): 'Machine translation and machine-aided translation' *Journal of Documentation* 34, 1978, 119-159 (Progress in Documentation)
- Kay, M. (1973): 'The MIND system' In: Rustin, R. (ed.) *Natural language processing* (New York, Algorithmics Pr., 1973), 155-188
- Kay, M. (1975): 'Automatic translation of natural language' In: Haugen, E. & Bloomfield, M. (eds.) *Language as a human problem* (Guildford, Lutterworth, 1975), 219-232
- Kulagina, O. S. et al. (1971): *Ob odnoi vozmozhnoi sisteme mashinnogo perevoda*. Moskva, Inst. Russkogo Yazyka AN SSSR, 1971 (Predvaritel'nye Publikatsii, Vyp. 21)
- Labov, W. (1975): 'Empirical foundations of linguistic theory' In: Austerlitz, R. (ed.) *The scope of American linguistics* (Lisse, de Ridder, 1975), 77-133 (Also publ. as: *What is a linguistic fact?* Lisse, de Ridder, 1975)
- Lamb, S. M. (1966): *Outline of stratificational grammar*. Washington, D.C., Georgetown Univ. Pr., 1966

- Lehmann, W. P. and Stachowitz, R. (1972-75): Development of German-English machine translation system. Final (annual) report(s). Austin, Univ. Texas, Linguistics Research Center, 1972(-1975)
- Mel'chuk, I. A. and Zholkovskii, A. K. (1970): 'Towards a functioning 'meaning-text' model of language' *Linguistics* 57, 1970, 10-47
- Popper, K. R. (1972): *Objective knowledge: an evolutionary approach*. Oxford, Clarendon Pr., 1972
- Roberts, A. H. and Zarechnak, M. (1974): 'Mechanical translation' In: *Current trends in linguistics, vol. 12: Linguistics and adjacent arts and sciences*, pt.4 (The Hague, Mouton, 1974), 2825-2868
- Sinaiko, H. W. and Klare, G. R. (1972): 'Further experiments in language translation: readability of computer translations' *ITL* (Review of Institute of Applied Linguistics, Louvain) 15, 1972, 1-29
- TAUM (1973): 'Le système de traduction automatique de l'Université de Montréal (TAUM)' *META: Journal des Traducteurs* 18, 1973, 227-289
- TAUM-Aviation (1977): TAUM Projet Aviation. Rapport d'étape, mai 1977. Montréal, 1977 (mimeo)
- Tesnière, L. (1959): *Eléments de syntaxe structurale*. Paris, Klincksieck, 1959
- Toma, L. et al. (1974): Some semantic considerations in Russian-English machine translation. Final report. Griffiss AFB, New York, Rome Air Development Center, 1974
- Toma, P. (1977): 'SYSTRAN as a multilingual machine translation system' In: Commission of the European Communities. *Overcoming the language barrier* (München, Vlg. Dokumentation, 1977), 569-581
- Vauquois, B. (1975): *La traduction automatique à Grenoble*. Paris, Dunod, 1975
- Weaver, W. (1955): 'Translation' In: Locke, W. N. and Booth, A. D. (eds.) *Machine translation of languages* (New York, Wiley, 1955), 15-23
- 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
- Winograd, T. (1972): *Understanding natural language*. Edinburgh, Univ. Pr., 1972
- Woods, W. A. (1970): 'Transition network grammars for natural language analysis' *Communications of the ACM* 13, 1970, 591-606
- Zholkovskii, A. K. and Mel'chuk, I. A. (1970) 'Sur la synthèse sémantique' *T.A. Informations* 1970 (2)