Chapter 3: Problems, methods, and strategies.

3. 1: Mechanical dictionaries

The creation of an automatic dictionary is the first and most obvious task of an MT system. Mechanical dictionaries were the central concern of all the earliest MT researchers and they are still crucial for the efficient operation of present MT systems. Like Artsrouni and Troyanskii, many early researchers tended to see the translation process almost exclusively in terms of consulting dictionaries for finding TL words equivalent to SL words. The resulting 'dictionary translations' presented the TL output in the same word sequence as the SL input, i.e. 'word-for-word' translations. They knew that this would not produce good translations; they expected the results to be very poor and in need of considerable editing. Before any 'word-for-word' translations had been seen, Reifler suggested a pre-editor and a post-editor (ch.2.4), and the unintelligibility of the results of Richens and Booth's attempts (ch.2.2) confirmed the need for, at the least, post-editing. Nevertheless, the ability of many readers to make some sense of these dictionary translations encouraged MT researchers to believe that with suitable modifications the 'word-for-word' approach could in the end produce reasonable output. As we have seen, Yngve considered that they were "surprisingly good" and worth taking as first approximations to be worked on.

The mechanisation of dictionary procedures posed problems of a technical nature. Research on MT began at a time when computers were limited in their storage capacities and slow in access times. There was much discussion of storage devices and mechanisms for improving access times. Booth (1955) and Stout (1954), for example, assessed the relative merits of paper tape, punched cards and magnetic tape as external storage means and the various possibilities for internal 'memory' storage, cathode-ray-tube dielectric stores, vacuum tubes, magnetic drums, photographic drums, etc. Since the external storage could only be searched serially, the most efficient method of dictionary lookup was to sort all the words of the SL text into alphabetical order and to match them one by one against the dictionary entries. Once found, entries could often be stored internally where faster access was possible. Various proposals were made for efficient searching of internal stores, including the sequencing of items by frequency, the 'binary cut' method first put forward by Booth (1955a), and the letter-tree approach of Lamb (ch.4.10).

A popular method for reducing dictionary size was the division of words into stems and endings. In languages like German and Russian it was obviously wasteful to include every inflected form of nouns and verbs. The familiar regularities of noun and verb paradigms encouraged researchers to investigate methods of morphological analysis to identify stems and endings. However, there are so many peculiarities and irregularities in the morphology of languages that procedures turned out to be more complex than expected; as a result, when larger storage mechanisms with fast access times became available many MT researchers went back to the older system of storing full forms in dictionaries.

Obviously, dictionaries cannot always include all the words occurring in SL texts. A problem for all MT systems is to establish acceptable methods for dealing with missing words; basically, there are two approaches, either to attempt some kind of analysis and translation, or to print out the original unlocated SL form. In both cases, there is a further problem with the rest of the sentence; whether to attempt an incomplete translation or to give up and produce no translation. In experimental MT systems it is obviously reasonable to admit failure, but in operational systems it is desirable, on the whole, to produce some kind of translation.

3. 2: Polysemy and semantics.

The most obvious deficiency of any word-for-word translation, whether mechanised or not, is that the order of words in the resulting TL text is more often wrong than correct. As we have seen, it was clear to Oswald and Fletcher (ch.2.4.1) that translation of German texts into English

demanded some kind of structural analysis of the German sentences. At the simplest level, such analysis may take into account morphological features, such as the endings of nouns, adjectives and verbs, or basic syntactic sequences, such as noun-adjective and subject-verb relations. As we shall see, it is possible to use this kind of information to devise procedures for rearrangement in basically 'word-for-word' systems. But, in order to go beyond the inherent limitations of the word-for-word approach, the analysis of syntactic structures must involve the identification of phrase and clause relationships. Methods of syntactic analysis will be the subject of the next section.

The second obvious problem is that there are rarely one-to-one correspondences in the vocabularies of natural languages. In most cases, a particular SL word may correspond to a number of different TL words, so that either the MT system prints out all the possibilities or it attempts to select the one which is most appropriate for the specific text in question. The first option was adopted by many systems, as we shall see, often as an intermediate stop-gap; the problem of selecting the right TL equivalent was left to the post-editor. Attempts to deal with the problem took a number of approaches.

The difficulty occurs usually because the SL word has what Weaver and many after him called 'multiple meanings'. Linguists distinguish between homonyms and polysemes; homonyms are words like *bank* which have two or more distinct and unrelated meanings ('geological feature' or 'financial institution'); polysemes are words like *face* which reflect different shades of meaning according to context. They distinguish also between homophones (words which sound the same but have different meanings) such as pear, *pair* and *pare*, and homographs (words which are spelled the same but have different meanings) such as *tear* ('crying' versus 'ripping'). Fortunately, the homophone problem is irrelevant since MT deals only with written texts. For practical purposes it is also immaterial whether the SL word is a homograph or a polyseme: the problem for MT is the same, the relevant meaning for the context must be identified and the appropriate TL form must be selected. Consequently, it is now common in MT research to refer to methods of 'homograph resolution', whether the words concerned are strictly homographs or not.

Sometimes the TL vocabulary makes finer sense distinctions than the SL. There are familiar examples in translating from English into French or German: the verb *know* may be conveyed by *savoir* or *connaître* in French and by *wissen* or *kennen* in German; likewise the English *river* may be either *rivière* or *fleuve* in French and either *Fluss* or *Strom* in German. In neither case can we say that the English words have more than one meaning; it is just that French and German make distinctions which English does not. Nevertheless, in the context of a MT system the problem of selecting the correct TL form is much the same as when the SL form is a genuine homograph or polyseme. MT systems do, however, differ according to whether this type of SL-TL difference is tackled at the same stage as SL homograph resolution or not (see ch.3.9 below on MT strategies).

The difficulties are further compounded in languages like English where many words may function as nouns, verbs or adjectives without any formal distinctions; e.g. *control* can be a verb or noun, green can be an adjective or a noun. The fact that there can be stress differences, e.g. between the verb *permit* and the noun permit, is of no assistance. For practical purposes these forms are also treated as homographs and much the same procedures for 'homograph resolution' are applied.

Various methods for tackling such SL-TL lexical differences have been proposed. One has already been mentioned, the identification of grammatical category either by morphological clues or by syntactic analysis. For example, the endings '-ed' and '-ing' generally indicate participial forms of English verbs (although they may be functioning as adjectives). Similarly, if in a two word sequence the first is definitely an adjective the second is probably a noun. Therefore, homographs which happen to belong to different syntactic categories may sometimes be distinguished in this way.

Another method is to reduce the incidence of homography in the MT dictionaries. The concept of the 'micro-glossary' was proposed (ch.2.4.3) not only to keep the size of dictionaries reasonably small but also to minimize problems of 'multiple meanings'. It was maintained, for

example, that the Russian *vid* was to be translated usually as *species* in biological contexts and not as *view*, shape or aspect. A micro-glossary for Russian-English translation in biology could, therefore, include just one of the English equivalents. In many cases the entry has to be the equivalent which is most often correct. In physics, for example, Russian izmenenie is usually equated with *change*; although in some contexts other translations may be better, the one which fits best most frequently should be selected.

The suggestion by Weaver was to examine the immediate context of a word. As we have seen, Kaplan concluded that a five word sequence was in general sufficient 'micro context' for disambiguation, i.e. for identifying the particular meaning of a polyseme. There are two ways in which immediate context can be implemented: one by expanding dictionary entries to include sequences of two or more words, i.e. phrases, the other by testing for the occurrence of specific words. For example, if the word *obrazovanie* is modified by *kristallicheskoe* then it is to be translated formation (rather than *education*); either the dictionary includes the whole phrase or the analysis procedure tests for the particular adjective. The dictionary solution obviously requires storage facilities of sufficient capacity, and it is also more appropriate when phrases are 'idiomatic', i.e. when the meaning (translation) of the phrase as a whole cannot be deduced (or constructed) from its individual words. Apart from familiar idioms such as *hold one's tongue, not move a finger, red herring* and *blue blood*, it could include verbal phrases such as *make away with, draw forth, look up* and *pass off*, and noun phrases such as *full speed, upper class* and *brute force*.

A more fundamental use of contextual information is the search for semantic features which are common to or prominent in the sentence or text as a whole, and to use this information to decide on the most fitting translation for SL words. This method involves the investigation of semantic 'invariants' or semantic regularities in vocabulary and texts, and necessarily goes far beyond the examination of lexical equivalents between languages. It involves, for example, the investigation of synonymy and paraphrase, of semantic 'universals' or 'primitive' elements (e.g. features such as 'human', 'animate', 'liquid', etc.), and of semantic relations within sentences and texts (e.g. agent-action, cause-effect, etc.)

Finally, the problem of polysemy may simply be avoided completely by insisting that texts input to a MT system be written in a regularized and normalized fashion. In other words, writers are encouraged not to be ambiguous, or rather not to include words and phrases which the MT system in use has difficulty in disambiguating. The obverse of this is to 'solve' polysemy by using a highly restricted form of TL as output, a kind of 'pidgin' language with its own idiosyncratic vocabulary usages. As we have seen, the first suggestion of this approach was made by Dodd (ch.2.4.3 above); the groups at the University of Washington and at Cambridge were particularly interested in MT pidgin and methods of improving output of this kind.

In theory, any of these methods can be used in any MT system; in practice, particular MT systems have emphasised one or two approaches, they have concentrated on exploiting their full potentialities and have generally neglected the alternatives. Concentration on the contextual and micro-glossary approaches was characteristic of the MT groups at Rand and Michigan. Concentration on the dictionary and lexicographic approaches was characteristic of the groups at Harvard, at the University of Washington and at IBM. Concentration on text semantics was pursued most strongly by the Milan group with its 'correlational analysis' approach and by the Cambridge group with its 'thesaurus' approach.

3. 3: Morphological analysis

In order to perform any kind of syntactic analysis the grammatical categories (noun, verb, adjective, adverb, etc.) of the words of sentences must be determined. The first step of analysis in any MT system is, however, the identification of the words in the SL text. This is relatively easy in English and most European languages, since words are separated by spaces in written text, but it is

not for example in languages such as Chinese and Japanese where there are no external markers of word boundaries.

Obviously, dictionary entries could indicate the grammatical categories ('word class' or 'part of speech') of all SL words. However, it was clearly unnecessary 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. 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 a first step towards the determination of grammatical categories, e.g. to continue the 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. As these examples demonstrate, however, many (perhaps most) endings are ambiguous, even in Russian, and the final establishment of the grammatical category of particular words in text takes place during syntactic analysis. Morphological analysis deals necessarily with regular paradigms; irregular forms, such as the conjugation of verbs such as *be* and *have*, and the plural forms of nouns such as *geese* and *analyses*, are generally dealt with by inclusion of the irregularities in full forms in the dictionary.

3. 4: 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. Examples are to be found in the 1954 Georgetown-IBM experiment (ch.4.3), and in the experiment by Panov and his colleagues shortly afterwards in the Soviet Union (ch.6.1). When there were differences of syntactic structure more complex than inversion, the solution was often the inclusion of phrases in the dictionary, i.e. rather like idiomatic expressions. This approach was expanded and refined as the 'lexicographic' approach of the University of Washington (ch.4.1).

Rearrangement rules may take into account fairly long sequences of grammatical categories, but they do not imply any analysis of syntactic structure, e.g. the identification of a noun phrase. The next step beyond the basic word-for-word approach is therefore the establishment of syntagmas, such as noun phrases (nouns and modifiers, compound nouns, etc., verbal complexes (e.g. auxiliaries and modals in conjunction with infinitives or participle forms), and coordinate structures. This level of analysis is to be seen in the later 'Georgetown system' (ch.4.3). Complete syntactic analysis involves the identification of relationships among phrases and clauses within sentences.

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. These are linear (or precedence) relations.

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. These are constituency relations.

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. These are hierarchical (or dominance) relations.

Included among the basic objectives of any method of syntactic analysis must be at least the resolution of homographs (by identification of grammatical categories, e.g. whether *watch* is a noun or a verb), the identification of sequences or structures which can be handled as units in SL-TL transfer, e.g. nouns and their associated adjectives.

Various models of syntactic structure and methods of parsing have been adopted in MT systems and are described in more detail in connection with particular MT projects in the following chapters. At this point, the main approaches will be outlined, illustrated in the most part 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 (parsing program) to identify sequential (linear) information was the Predictive Syntactic Analyzer developed at the National Bureau of Standards and at Harvard University (ch.4.8 and 4.9) The premise 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 determinable 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 grammatical category to which the word had been assigned. At the beginning of the analysis certain sentence types were predicted in terms of sequences of grammatical categories. Examination of each word was in two stages: first to test whether its category '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.1). The analysis of a sentence was completed if a terminal state has been reached and all categories have been accounted for. Initially, only the single most probable path through the series of predictions was taken during parsing, but in later models all possible predictions were pursued. The method did not in principle need to recognise phrase structures or dependency relations, although these could be derived from the identification of specific category sequences. (This 'building' facility was employed in a later development, the ATN parser, ch.9.13)

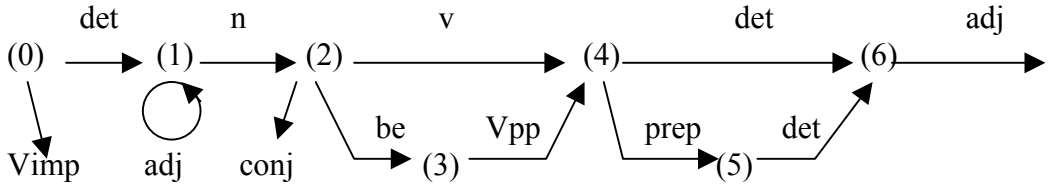

Fig.1. Finite state grammar

The second approach, analysis of dependency relations, is based on the identification of 'governors', e.g. the 'head' noun in a noun phrase, and their dependants or 'modifiers', e.g. adjectives. The governor of the sentence as a whole is generally taken to be the finite verb since this specifies the number and nature of dependent nouns (fig.2). A verb such as *buy*, for example, can have four dependants (purchaser, object purchased, price, seller) – a concept referred to as 'valency': a transitive verb such as *see* has a valency of two, an intransitive such as *go* has a valency of one, etc. (If the valency relationships are themselves specified as, e.g. 'agency', 'instrumentality', then we have a 'case frame' analysis, ch.9.16 below.)

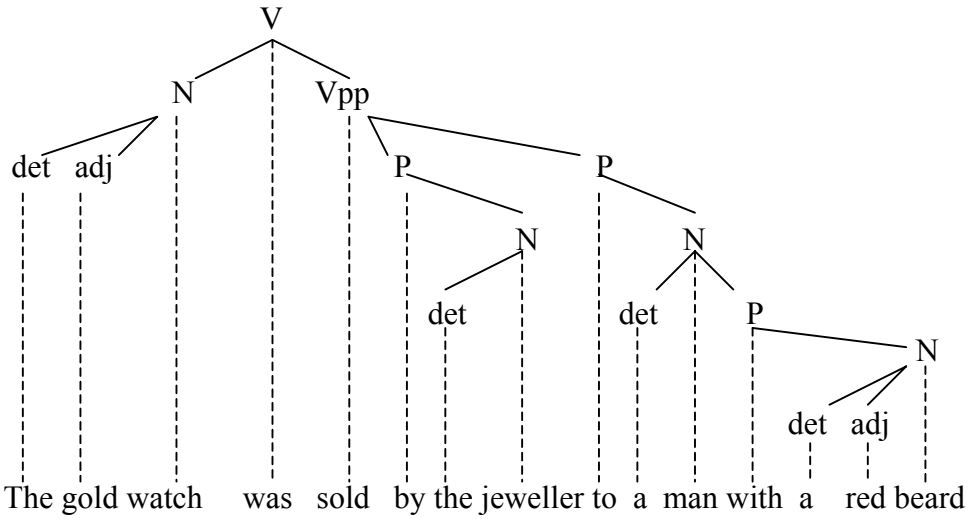

Fig.2. Dependency structure analysis

The parsing of dependency structure can operate either top-down (identification first of governors and then dependants) or bottom-up (determination of governors by a process of substitution). The top-down approach was most common, and can be illustrated by Garvin's fulcrum parser (ch.4.6): in a series of passes the algorithm identified 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. An example of bottom-up parsing may be seen in the 'set-theoretic' approach of Kulagina (ch.6.2)

The third approach, that of phrase structure analysis (fig.3), provides labels for constituent groups in sentences: noun phrase (NP), verb phrase (VP), prepositional phrase (PP), etc. The phrase structure approach is associated most closely in the early period of MT research with the MIT project (ch.4.7). Parsing can be either bottom-up or top-down. In the former, structures are built up in a series of analyses from immediate constituents, e.g. first noun phrases, then prepositional structures, then verb relationships and finally the sentence structure as a whole. In top-down parsing, the algorithm seeks the fulfilment of expected constituents NP, VP, etc. by appropriate sets and sequences of grammatical categories. The bottom-up parsing strategy was the most common approach in the early MT system, but at MIT some investigation was made into the top-down strategy (‘analysis by synthesis') In systems since the mid-1960's this strategy is now probably more common.

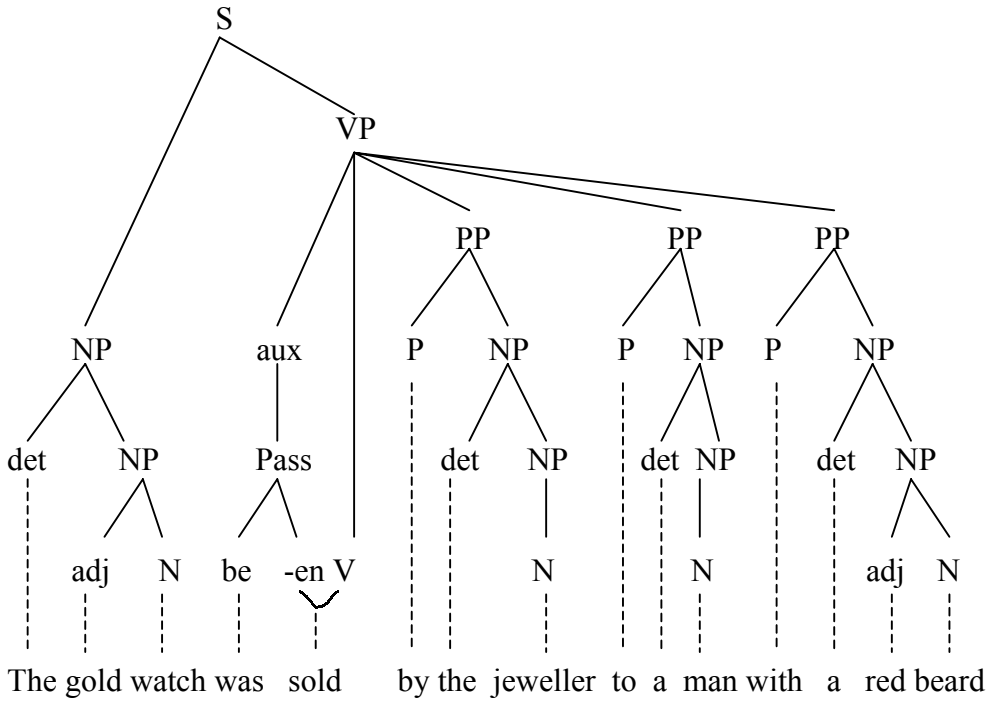

Fig.3. Phrase structure analysis

It may be noted that categorial grammar developed by Bar-Hillel (1960, app. II), which was one of the first attempts at formal syntax, is a version of constituency grammar. In a categorial grammar, there are just two fundamental categories, sentence s and nominal *n*; the other grammatical categories (verb, adjective, adverb, etc.) are defined in terms of their potentiality to combine with one another or with one of the fundamental categories in constituent structures. Thus a transitive verb is defined as n*s* because it combines with a nominal (to its left) to form sentences; and an adjective is defined as *n/n* because in combination with a nominal n to its right it forms a (higher-order) nominal *n*. In other words, the category symbols themselves define how they are to combine with other categories. Combination operates by two simple 'cancellation' rules: $x/y,\ y \rightarrow x$, and $y,\ y\backslash x \rightarrow x$.

3. 5: Formal syntax and transformational grammar

Research in MT helped to stimulate much interest in formal linguistics. An early result of this mathematization of syntax and linguistic theory was the demonstration that all phrase structure and dependency grammars are formally (i.e. mathematically) equivalent and that since they can be implemented on push-down automata, they are equivalent also to the so-called finite state grammars (Gross & Lentin 1967). All these grammars belong to the class of 'context-free' grammars. A context-free grammar consists of a set of rewriting rules (or production rules) of the form $A \rightarrow a$, where *A* belongs to a set of 'non-terminal' symbols and *a* is a string of non-terminal and/or terminal symbols. Non-terminal symbols are grammatical categories (S, NP, VP, N, Adj, etc.) and terminal symbols are lexical items of the language. Context-free grammars are important not only as the basis for formal grammars of natural languages but as the basis for computer programming, since the standard algorithmic methods used in compilers rely on finding only context-free structures in programming languages.

However, Noam Chomsky (1957) 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. Context-free grammars are unable, for example, to relate different structures having the same functional relationships, e.g.

where discontinuous constituents are involved: *He looked up the address* and *He looked the address up*; or where there are differences of voice, e.g. the active: *The jeweller sold the watch to the man yesterday* and the passive: *Yesterday the man was sold the watch by the jeweller*. Chomsky proposed a transformational-generative model which derived 'surface' phrase structures from '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 appears as the 'deep' logical object (fig.4). Deep structures are generated from an initial symbol S by 'context-sensitive' rewriting rules. An essential feature of the Chomskyan model is that syntactic structures are generated top-down from initial symbol S, to 'deep' structure tree and then by transformational rules to 'surface' structure trees. In the case of a coordinate phrase such as gold watch and chain the base 'deep' structure would make explicit the fact that both watch and chain are gold (fig.5a). To produce the elliptical 'surface' form (fig.5b) a transformation rule would delete the repeated adjective. The model is not intended to provide the basis for a recognition grammar (e.g. a parser), but only to define mathematically the set of well-formed sentences, and to assign "a structural description indicating how the sentence is understood by the ideal speaker-hearer" (Chomsky 1965: 5). The implications of this approach became clearer when researchers attempted to develop 'transformational parsers' (ch.9.11)

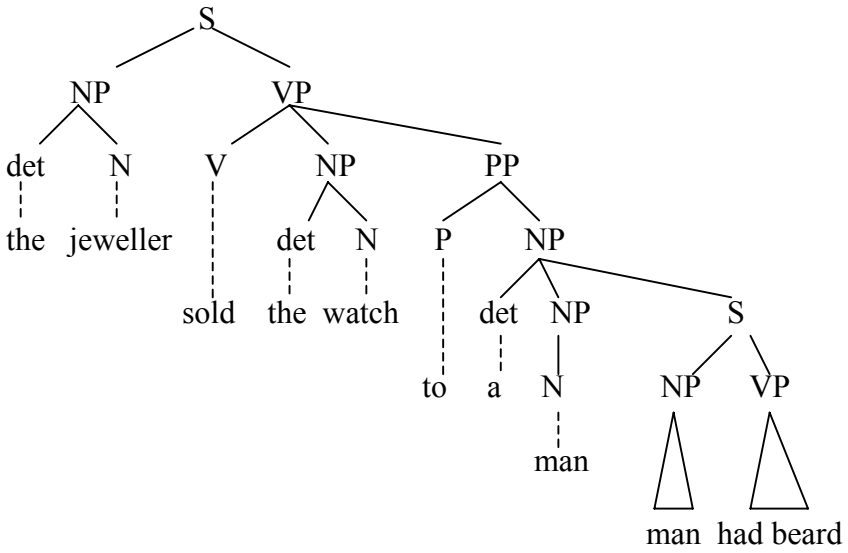

Fig.4. 'Deep' structure analysis

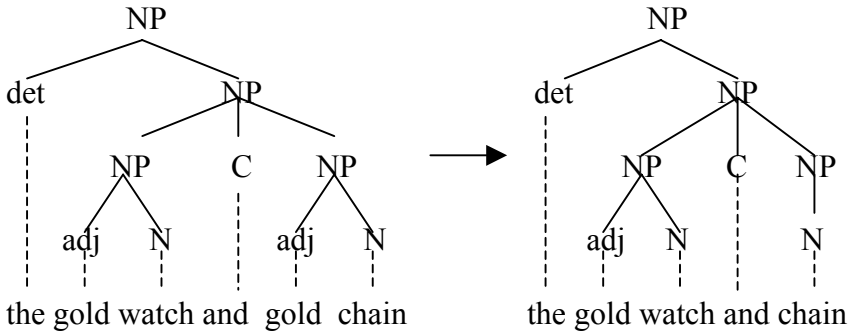

Fig.5. Transformational rule (loss of phrase structure relationship)

Chomsky's notion of transformational rules derived formally from the work of Zellig Harris (1957). Harris' concern was the development of a symbolism for representing structural relationships. Grammatical categories were established primarily on the basis of distributional analysis. Thus, the subject of a sentence can be a (single) noun (*The man...*), a clause (*His leaving home...*), a gerundive (*The barking of dogs...*), an infinitive clause (*To go there...*), etc. In order to function as subjects, clauses have to undergo transformations from 'kernel' (atomic sentence-like) forms: e.g. *He left home* → *His leaving home*, *Dogs bark* → The *barking of dogs*. For Harris, transformations were a descriptive mechanism for relating surface structures, while in Chomsky's model, transformational rules derive surface structures from 'deeper' structures. By the mid-1960's (Chomsky 1965) an additional requirement of transformational rules was that they should be 'meaning-preserving', i.e. from a 'deep' structure should be generated semantically equivalent surface structures. Although Chomsky's syntactic theory has undoubtedly had most influence, the formalisation of transformations by Harris had considerable impact in MT research, particularly in the representation of SL-TL structural transfer rules.

3. 6: Syntactic ambiguity and discourse relations.

Although the identification of grammatical categories and of sentence structures is clearly important in linguistic analysis, there are inherent limitations in syntactic analysis which were recognised before even efficient parsers had been developed. 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 + NP1 + P + NP2 + P + NP3 must have parsings which relate NP2 and V, NP2 and NP1, NP3 and V, NP3 and NP2 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 been applied in many MT systems since the mid-1960's (as the following descriptions of systems will show). However, semantic features 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 most frequent occasions on which recourse is made to 'real world' knowledge involve the reference of pronouns. Examples are the two sentence pairs:

The men murdered the women. They were caught three days later.

The men murdered the women. They were buried three days later.

The correct attribution of the pronoun *they* to the men in the first pair and to the women in the second depends entirely on our knowledge that only dead people are buried, that murder implies death, that murder is a criminal act, and that criminals ought to be apprehended. This knowledge is non-linguistic, but it has linguistic implications in, for example, translation of these sentences into French where a choice of *ils* or *elles* must be made.

Of course, it is not only in cases of syntactic ambiguity that we use 'real world' knowledge to help in understanding text. Homographs can, as indicated earlier, be resolved by identification of grammatical categories, e.g. whether *watch* is a noun or a verb. However, the resolution of some homographs 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. It was from such instances that Bar-Hillel was to argue in an influential paper (Bar-Hillel (1960) that fully automatic translation of a high quality was never going to be feasible (ch.8.3 below). In practice this type of problem can be lessened if texts for translation are restricted to a more or less narrow scientific field, and so dictionaries and grammars can concentrate on a specific 'sublanguage' (and this was the argument for 'micro-glossaries'). Nevertheless, similar examples recur regularly, and the argument that MT requires 'language understanding' based on encyclopaedic knowledge and complicated inference procedures has convinced many researchers that the only way forward is the development of 'interactive' and Artificial Intelligence approaches to MT (ch.15 and 17)

In general, semantic analysis has developed, by and large, as an adjunct of syntactic analysis in MT systems. (Exceptions are those MT systems with an explicitly semantic orientation, cf.9.17 below) In most 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 (as in a case frame specification, ch.9.16 below). The 'shepherd's staff' sense of *crook* is thus excluded. In many systems semantic features have been assigned as 'selection restrictions' in an ad hoc manner, as the demands of the analysis of a particular group of lexical items seem to require them, and also somewhat too 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. One answer suggested has been to make such 'selection restrictions' define not obligatory features but preferences (ch.15.1)

True semantic analysis should include some decomposition of lexical items according a set of semantic 'primitives' or putative 'universals'. Only by such means is it possible to derive common semantic representations for a pair of sentences such as *The teacher paid no attention to the pupil* and *The pupil was ignored by the teacher*. In general, the majority of MT systems have avoided or held back from the intricacies and complexities and no doubt pitfalls of this kind of semantics. It is found therefore only in those MT groups which have investigated interlinguas (e.g. the Cambridge group, ch.5.2, and the Soviet group around Mel'chuk, ch.10.2), and in some of those recent (since mid-1970's) groups with an interest in AI methods (ch.15)

3. 7: Sentences and texts

The difficulties with pronominal reference described above stem also from the exclusive concentration of syntax-based analysis on sentences. The need for text-based analysis can be illustrated by the following two German sentences:

In der Strasse sahen wir einen Polizist, der einem Mann nachlief. Dem Polizist folgte ein grosser Hund.

Translation into English sentence by sentence would normally retain the active verb forms producing:

In the street we saw a policeman running after a man. A large dog followed the policeman.

Text cohesion would be improved if the second sentence were passivized as:

The policeman was followed by a large dog.

This inversion requires that a MT system adheres as far as possible to the information structure of the original, i.e. in this case retains the 'policeman' as the head (or topic) of the sentence. The

problems of topicalisation and text cohesion are of course far more complex than this example. Scarcely any MT projects have even considered how they might be tackled.

3. 8: Transfer and synthesis.

The production of output text in the target language (TL) is based on the information provided from dictionaries and from the results of analysis. In general the synthesis of TL sentences is less complex than the analysis of SL input. The process involves nearly always the derivation of correct morphological forms for TL words (unless dictionaries contain only full TL forms). Thus, for example, TL synthesis must produce the right forms of verbs, e.g. for English simple past forms it is not a matter of just adding *-ed* as in *picked* (from *pick*), since sometimes endings must be deleted or altered as in *lived* (not: *liveed*) and *tried* (not: *tryed*), etc. Irregular forms are generally handled by the dictionary (e.g. *went* would be coded directly as the past form of *go*).

If analysis has included the establishment of syntactic structure (e.g. a phrase structure) then synthesis must convert this structure into an appropriate TL structure and produce a linear representation, i.e. it must invert the analysis process in some way. However, it should be stressed that inversion does not imply that the rules devised for the analysis of structures for a particular language (as SL) can be simply reversed to obtain rules for synthesis of that language (as TL).

At some point in many systems (the exceptions being interlingual systems, cf. next section), the syntactic structures of SL texts are transformed into TL structures. Whether such transformations apply to only short segments (as in word-for-word systems) or to whole sentences, the process involves the specification of transformation rules. For example, a rule for changing a German construction with final past participle (*Er hat das Buch gestern gelesen*) into an English construction with a simple past form (*He read the book yesterday*) might be in this form:

NP + aux + ... + Vpp --> NP + Vpst + ...

Clearly, such transformation rules have much in common with the transformation rules which Harris devised for relating structures within the same language.

3. 9: System designs and strategies

In broad terms, there have been three types of overall strategy adopted in MT systems.

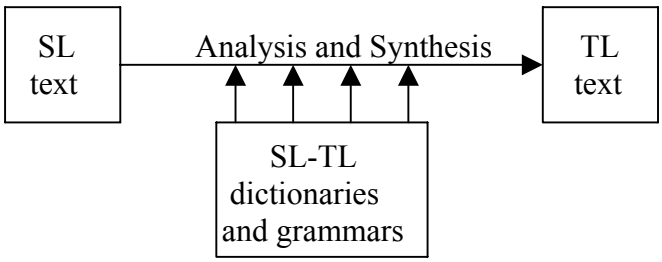

Fig.6. 'Direct translation' system

The first approach is the 'direct translation' approach (fig.6). Systems are designed in all details specifically for one particular pair of languages. The basic assumption is 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 is sufficiently close to an acceptable sequence of TL words, then there is no need to identify the syntactic structure of the SL text. The majority of MT systems of the 1950's and 1960's were based on this approach. They differed in the amount of analysis and/or restructuring incorporated. There was none at all in the straight 'dictionary translation' experiment of Richens and Booth (ch.2.2); there was just a minimum of local restructuring in the 'word-for-word' systems of the University of Washington and IBM (ch.4.1 and

4.2); there was partial analysis of SL structure in the Georgetown system (ch.4.3); and there was full sentence analysis in the systems at Ramo-Wooldridge, Harvard, and Wayne State University (ch.4.6, 4.9, 4.12). A primary characteristic of 'direct translation' systems of the earlier period was that no clear distinctions were made between stages of SL analysis and TL synthesis (cf. particularly the account of the Georgetown system below). In more recent (post-1970) examples of 'direct' systems there is a greater degree of 'modular' structure in the systems (cf.ch.9.10 below).

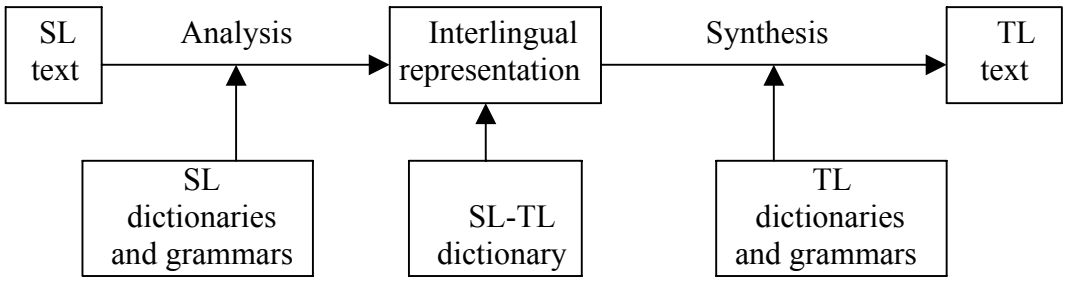

Fig.7. 'Interlingual' system

The second basic MT strategy is the 'interlingual' approach, which assumes that it is possible to convert SL texts into semantico-syntactic representations common to more than one language. From such 'interlingual' representations texts would be generated into other languages (fig.7). In such systems translation from SL to TL is in two distinct and independent 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. Procedures for SL analysis are intended to be SL-specific and not devised for any particular TL in the system; likewise, TL synthesis is intended to be TL-specific. Interlingual systems differ in their conceptions of an interlingual language: a 'logical' artificial language, or a 'natural' auxiliary language such as Esperanto; a set of semantic primitives common to all languages, or a 'universal' vocabulary, etc. Interlingual MT projects have also differed according to the emphasis on lexical (semantic) aspects and on syntactic aspects. Some concentrated on the construction of interlingual lexica (e.g. the Cambridge and the Leningrad groups); others have concentrated on interlingual 'syntax' (e.g. the Grenoble and Texas groups).

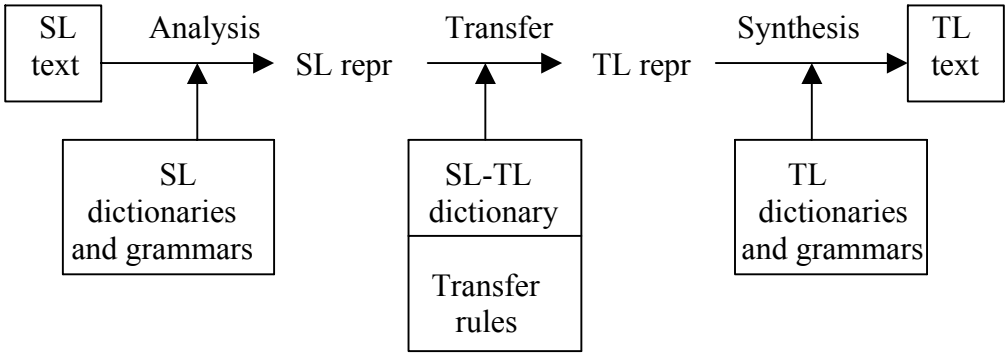

Fig.8. 'Transfer' system

The third approach to overall MT strategy is the 'transfer' approach (fig.8). Rather than operating in two stages through a single 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 'transfer' representations, the second converts these into TL 'transfer' 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. Differences between languages of the know-*savoir/connaître* type (section 3.2 above) would be handled during transfer. In English analysis, *know* is treated as unambiguous, there is no need to determine which kind of 'knowing' is involved. Whereas the 'interlingual' approach would require such analysis, the 'transfer' approach does not; problems of mismatch between SL and TL lexical ranges are resolved in the transfer component. Systems differ according to the 'depth' of analysis and the abstractness of SL and TL transfer representations. In the earliest systems (for which the MIT system (ch.4.7) is the exemplar), analysis went no further than 'surface' syntactic structures, with therefore structural transfer taking place at this depth of abstraction. Later (post-1970) transfer systems have taken analysis to 'deep' semantico-syntactic structures (of various kinds), with correspondingly more abstract transfer representations and transfer rules (cf. ch.13 on TAUM, GETA, etc.)

The basic difference between these two 'indirect' approaches and the (generally earlier) 'direct' approach lies in the configuration of dictionary and grammar data. In 'direct' systems the main component is a single SL-TL bilingual dictionary incorporating not only information on lexical equivalents but also all data necessary for morphological and syntactic analysis, transfer and synthesis. In 'indirect' systems, this information is dispersed among separate SL and TL dictionaries, separate SL and TL grammars, and either the interlingua vocabulary and syntax, or the SL-TL 'transfer' dictionary (of lexical equivalences) and a 'grammar' of SL-TL structure transfer rules (see figs. 6 to 8).

3. 10: Perspectives and influences.

While the classification of MT systems in terms of basic strategy is a convenient descriptive device and will be employed in the grouping of system descriptions in later chapters, it has not been the most prominent perspective for MT researchers, particularly in the 1950's and 1960's. For this period, the most important distinctions were between the engineering and the 'perfectionist' approaches, between the empiricist and other methodologies, and between the syntax orientation and various lexical and word-centred approaches.

The most immediate point of dispute was between those groups who agreed with Dostert and Booth on the importance of developing operational systems as quickly as possible (ch.2.4.3) and those who argued for more fundamental research before such attempts. The engineering approach held basically that all systems can be improved and that the poor quality early word-for-word systems represent a good starting point. There were differences between what Garvin (1967) dubbed the 'brute force' approach, which assumed that the basic need was larger storage capacity (e.g. the IBM solution, ch.4.2), and the engineering approach which believed that algorithmic improvements based on reliable methods of (linguistic) analysis could lead to better quality (the Georgetown, ITMVT, Ramo-Wooldridge projects). The 'perfectionists' included all those groups which concentrated on basic linguistic research with 'high quality' systems as the objective (MIT, Harvard, Rand, Berkeley, MIAN, etc.) The latter differed considerably in both theories and methods. Disputes between the 'perfectionists' and the 'engineers' recurred frequently until the mid-1960's (cf.ch.8.2).

On questions of methodology the main point of difference concerned the 'empiricist' approach, exemplified by the RAND group (ch.4.4). The approach emphasised the need to base procedures on actual linguistic data; it was distrustful of existing grammars and dictionaries; it believed it was necessary to establish from scratch the data required and to use the computer as an aid for gathering data. The approach stressed statistical and distributional analyses of texts, and a 'cyclic' method of system development: i.e. routines devised for one corpus were tested on another, improved, tested on a third corpus, improved again, and so forth. The empirical approach was in

fact fully in accord with the dominant linguistic methodology of the 1940's and 1950's in the United States, the descriptivist and structuralist 'tradition' associated particularly with Leonard Bloomfield (1933). The descriptivists adopted the behaviourist and positivistic method which insisted that only interpersonally observed phenomena should be considered 'scientific' data, and which rejected introspections and intuitions. They distrusted theorising, stressed data collection, and concentrated on methods of discovery and analysis. Traditional grammars were suspect: Charles Fries (1952), for example, undertook a distributional analysis of telephone conversations which resulted in new grammatical categories for English. Most descriptivists worked, however, on phonetics and phonology. Only in the mid-1950's did some descriptivists such as Zellig Harris start work on syntax. It was therefore not surprising that the empiricists regarded their research within MT as extending the range of descriptive linguistics.

The 'empiricist' emphasis on probabilistic and statistical methods, however, has perhaps a different origin. It is likely to be the considerable influence of the statistical theory of communication associated with Claude Shannon, i.e. 'information theory', and to which Warren Weaver made a substantial contribution (Shannon & Weaver 1949). The theory had great impact on the anti-metaphysical inclinations of most American linguists, since it seemed to provide a basis for developing mechanical methods for 'discovering' grammars. It may be noted that when Yngve first presented his ideas on 'syntactic transfer' Yngve (1957), he related his tripartite model to the information-theoretic triple of sender, channel and receiver.

A third area of difference in early MT groups was the question of what should be taken as the central unit of language. The majority assumed the centrality of the sentence; their approach was sentence-oriented (as was and still is, in essence, that of most linguists and logicians), and so there was an emphasis on syntactic relations and problems. A minority upheld the centrality of the word. They emphasised lexical and semantic relations and problems. They included the 'lexicographic' approach of Reifler and King, the 'thesaural' approach of the Cambridge group, the 'word-centred' theories of Lamb at Berkeley, and the dictionary-centred aspects of Mel'chuk's 'meaning-text' approach. It should be stressed that these are 'differences' only of general orientation; the 'syntax-oriented' groups did not neglect lexical and semantic issues, and the 'lexis-oriented' groups did not by any means neglect syntax. Indeed, in the case of Lamb and Mel'chuk it is very much an open question whether their models can be said to be oriented one way or the other.

In the descriptions above of various aspects of MT system design and methods of analysis, it may well have been implied, at a number of points, that language systems are intrinsically multi-levelled; that is to say, that linguistic description is necessarily couched in terms of phonetics, phonology, morphology (word formation), syntax, and semantics; and furthermore, that analysis proceeds through each of these levels in turn: first morphological analysis, then syntactic analysis, then semantic analysis. (The most extensive 'stratificationalist' models were in fact developed within the MT context, by Lamb and Mel'chuk.) Although undoubtedly a 'stratal' view of language systems is dominant in linguistics and has been since the time of Saussure, the founder of modern (structuralist) linguistics, it has not been the conception of some MT project teams. Indeed, many (particularly in the earliest period) would have rejected such a stratal view of language both for being too rigid and for not conforming to reality. For them, all aspects of language (lexical, semantic, structural) interact inextricably in all linguistic phenomena. There is no doubt that among the most linguistics-oriented MT groups there has been sometimes an excessive rigidity in the application of the stratal approach to analysis (e.g. in parsing systems); and it has led to failures of various kinds (cf.ch.9.2 and 9.12). Nevertheless, the basic validity of the approach has not been disproved, and most modern (linguistics-oriented) MT systems retain this basic conception (e.g. the GETA and Eurotra systems).

3. 11: MT research in the period 1956-66.

In the years immediately after the 1954 Georgetown-IBM demonstration, MT research began to receive massive funding in the United States, and to a lesser extent elsewhere. The level of support from military and intelligence sources can be explained partly by the prevailing political climate and by US fears of being overtaken by the Soviet Union in technology an science. However, it can be explained also as a reflection of the often exaggerated visions of computers as 'thinking machines'. Of all the possible tasks which these 'electronic brains' could be expected to do in the future, one of the most useful and practicable appeared to be translation. MT research became in consequence the pioneer field for work in 'artificial intelligence' (although the term was not to be coined until 1956 (McCorduck 1979)).

Much of the research in the 1950's and 1960's which was undertaken by MT groups would now be rightly regarded as belonging to other disciplines and fields. A number of projects, for example, were equally interested in problems of information retrieval and automatic indexing and abstracting, which were seen as closely linked to MT in so far as they involved complex linguistic analysis.

MT projects made major contributions to linguistic theory, principally formal grammar and mathematical linguistics. The work on syntactic analysis and parsing stimulated interest in the mathematical foundations of 'context-free' grammars (sect.3.4 above) Theoretical linguistics was the focus of much of the research at MIT, where a number of prominent linguists and transformational grammarians (beginning with Chomsky himself) were engaged in basic linguistic research. Elsewhere, at Rand, the foundations of dependency grammar were elaborated by Hays (1964) on earlier work by Tesnière (1959), and at Berkeley, Lamb developed his theory of stratificational grammar (Lamb 1966). Other centres with substantial contributions to linguistic research were Harvard, Texas, and MIAN in the Soviet Union.

Above all, MT was for many years the focus for all research in what is now called computational linguistics. Much of the activity at RAND was concerned with general applications of the computer to linguistic analysis and to linguistic data processing. From the mid-1960's MT became a peripheral activity at RAND, as indeed it did at Harvard and MIT. Computational linguistics grew from MT research and the two were regularly equated. The original name of the Association for Computational Linguistics was the Association for Machine Translation and Computational Linguistics (ch.8.2). A further indication of close links is the fact that the University of Pennsylvania group under Zellig Harris was regularly included in accounts of MT research during the 1960's, e.g. in the surveys by Bar-Hillel (1960) and Josselson (1971). Yet it is quite clear even from contemporary reports that Harris was engaged on basic syntactic research of English and had no intention of developing an experimental MT system (even in theory).

To a significant degree, in the 1950's and 1960's MT was the current most popular 'bandwaggon'; a fact which contributed to its later downfall (ch.8.9) The major sponsors in the US were the National Science Foundation, the Central Intelligence Agency, and the US military forces. For obvious reasons, their main interest was in Russian-English translation and this was the predominant language-pair studied by MT groups during this period. In the Soviet Union the main concentration was on English-Russian, but in general the MT research effort was more diversified than in the US. In the course of time, many countries had MT research activities of some kind; the major centres outside the US and the Soviet Union were Great Britain, Italy, and France.

In the following chapters the research groups will be discussed individually. The next chapter deals with the US projects, first projects which eventually produced operational systems: University of Washington, IBM, Georgetown (4.1-3), then the 'empiricist' projects at RAND, Michigan and the associated Ramo-Wooldridge project (4.4-6), then the more 'theoretical' groups at MIT, Harvard, Berkeley and Texas (4.7, 4.9, 4.10, 4.11). Also treated are the NBS and Wayne State and other 'minor' US projects (4.8, 4.12, 4.13). This long chapter is followed by one devoted to the groups in the United Kingdom, Birkbeck College (5.1), Cambridge (5.2) and the National

Physical Laboratory (5.4); to the Milan group (5.3); and to the French, and other Western European groups (5.5-8). The next chapter covers research in the Soviet Union and in Eastern Europe. Finally, chapter 7 deals with projects and research groups in Japan (7.1), China (7.2) and Mexico (7.3).

Major information sources for this period of MT research (apart from the accounts of individual projects) are the important surveys by Bar-Hillel (1960), Delavenay (1960), Josselson (1971), Mounin (1964), Pendergraft (1967), primarily for US research; and by Harper (1963), Ljudskanov (1972), Mel'chuk (1963), and Mukhin (1963) for Soviet research. Also invaluable are news items in the journals *Mechanical Translation* (1954-65), and *Traduction Automatique* (1960-64), and the project descriptions in the National Science Foundation's biannual *Current Research and Development in Scientific Documentation* (1957-66); these sources are abbreviated *MT*, TA and *CRDSD* respectively in the following chapters. Representative collections of articles are to be found in conference proceedings (Edmundson 1961, NATO 1966, NPL 1962) and in collections by Booth (1967) and Bruderer (1982). Bibliographic sources are the major survey articles and Delavenay (1960a), Van Hoof (1973: 464-504), Kent (1960a), Klein (1971) and, in particular, the critical bibliography by Mel'chuk & Ravich (1967).