[From: *Information Processing 1962*: proceedings of IFIP 62, Munich, 27 August – 1 Sept. 1962 (Amsterdam: North-Holland, 1963)]

THE USE OF COMPUTERS IN RESEARCH ON MACHINE TRANSLATION

OLGA S. KOULAGINA

Mathematical Institute of the Academy of Sciences of the USSR, Moscow, USSR

Read by A. A. DORODNICYN *(USSR)*

1. INTRODUCTION

Work on machine translation has been carried on in various countries for several years. During this time, a number of research groups have carried out machine translation of more or less extensive texts, from one language into another. In some cases, these were experimental translations of a relatively small number of sentences, in order to test the proposed algorithm; in other cases, long texts have been translated, but as a rule, with simplified algorithms and with reduced requirements as to the quality of translation. At present it is clear that the construction of algorithms by which computers can produce high-quality translation is quite feasible; but in order that these algorithms should have the necessary potentialities, and the computers be effectively used, it is necessary to solve a number of different theoretical problems. In order to achieve success in machine translation, it is essential to formulate the problems accurately, and to systematize both the problems and the difficulties arising in their solution. This is especially important if we consider not only the achievement of practical results in machine translation, but also the advance towards the wider use of computers to assist man in the various problems which involve the handling of verbal information, and which include MT as a special case.

2. PROBLEMS OF DESCRIPTION

Theoretical problems arising in MT can be divided into those of description and those of construction.

Problems of description are those related to the construction of systems for describing the languages to be translated, and the translation algorithms. The system for the description of a certain language must contain, on the one hand, a set of features by which the words of the language can be characterized, and on the other hand, the description of types of possible combinations of words that have certain features (i.e. grammatical configurations). Unfortunately there are, as yet, no such sufficiently complete, formalized, and precise grammars of the different languages. Usually, the description of a language is made in parallel with the construction of, and often in conformity with, the translation algorithm. For the study of the general properties of such systems, mathematical models of languages are useful. These models differ according to the approach to the language (analytic or synthetic), and according to the mathematical apparatus employed (theory of sets, theory of automata, theory of lattices, theory of graphs, etc.). One of the most important problems in the construction of a model is that of its interpretation, i.e., the relation between the model and the language.

The problem of the relation between the model and the language is especially difficult in the case of analytic models. A synthetic model can be evaluated by its results, and so experiments on the realization of such models, i.e. on the synthesis of sentences on the computer, are of extreme interest. The problem of the equivalence of models and, generally, the relation between models of different kinds, have been studied very little so far, though both the correspondence between model and language, and the relations between models, are very interesting. It appears reasonable to look for the following kind of relation between analytic and synthetic models; namely, that it should be possible to analyse any phrase generated by a given synthetic model within the given analytic model.

To problems of description there belong, as has been pointed out, not only systems for the description of languages, but also systems for the description of translation algorithms (algorithms of analysis and synthesis for these languages). At present, several such systems are known, for instance, the COMIT system developed by Yngve [1]), or the operator systems (Koulagina [2]), Meltchouk [3])). Each of these has its merits and demerits. The drawbacks are mostly due to the complexity of presentation, in the accepted form, of new cases of the transformation of information which had not been explicitly taken into account when developing the system for describing the algorithm. Owing to the universality of the suggested forms of description, the presentation of new transformations is always possible, though sometimes it turns out to be too cumbersome. In order to overcome these drawbacks, one should either seek ways of creating a general system which would include the possibilities of all other systems, or (what seems to be more reasonable) use different forms of presentation for different types of algorithm. Attempts to create a unique form of presentation for various algorithms would tend to lead to extreme complication and would make it difficult to use. Therefore, the development of systems for the description of translation algorithms is closely connected to the problem of typification and standardization of translation rules, of choosing appropriate types of algorithm, and of developing a general form of presentation, independent of the languages being handled.

Interesting work directed to the development of a standard form for an algorithm of morphological analysis, has been carried out by Meltchouk [4]). He has developed a general form of such an algorithm for a group of languages. It is possible to obtain an algorithm of morphological analysis for a certain concrete language from this general system, in accordance with the properties of the language in question. A questionnaire has been compiled concerning the properties of the language and, in accordance with the answers to the questionnaire, certain parts of the general algorithm are included in the algorithm for the given language.. Tables of affixes with their characteristics are specific for each language, while the rules for operating with these tables are common to all languages.

3. Problems Of Construction

The indication of certain types of translation algorithm, the development of a general form, and a system of formal notation, are essential to the solution of problems of construction.

To this group of problems there belong the problems of the construction of translation algorithms, and their realization by programs. These problems can be solved, either by starting directly from the source and target languages, or by considering their method of description, taking into account the ways of describing translation algorithms.

The standardization of translation rules, and the typification of translation algorithms, allows us to pass on to the automatic programming. The programming of translation algorithms requires much expenditure of effort, and the testing of a new translation algorithm takes a lot of time. With the standardization of translation rules, the programs of translation turn out to be compiled of parts which have similar structures. The presentation of an algorithm in the form of a series of standard operators, written in general in special language, allows us to entrust the programming to the computer itself, by means of a special programming program. One could go still further in passing over to the computer the greater and more complicated part of the work.

In the case where the problem of the construction of the translation algorithm is formulated as:

> Obtain the algorithm of a certain type, which can be written in a certain form, and which must handle information written in a certain form,

it is possible to arrive at a problem of formalization and standardization of the very process of construction of the algorithm.

4. Syntactic Analysis

In the Mathematical Institute of the Academy of Sciences of the USSR, research has already been carried out on the use of computers for the construction of algorithms for syntactic analysis of texts. By syntactic analysis we mean a stage of analysis where groups of words are investigated in order to find the connections between them; this is in contradistinction to morphological analysis, in the course of which we find all the information concerning the word, that can be obtained without reference to the context. The work of Meltchouk [5]) has shown that it is convenient to carry out syntactic analysis with a so-called *configuration table.* To establish the connection between the words of any sentence is equivalent to finding for each word, in accordance with certain rules, the governing word; in some cases a word can govern itself, i.e. be self-governing. The same word may in different cases, be governed by different words, and the task of the algorithm is to find the governing word for each word in each sentence. It should be borne in mind, however, that in a given sentence there can be several words, which according to their characteristics could govern the given word. In accordance with the above, the analysis algorithm must contain rules for finding governing words. Therefore, the algorithm of syntactic analysis (in its most simplified form) consists of rules of the following nature:

> If a word has given characteristics, then it can be governed by a word with certain characteristics in a given position with respect to it, if some specific additional conditions are satisfied; then the connection between them is of a certain type.

In this rule, the general form being fixed, the characteristics of the governing and the governed words, their respective order in the text, the additional conditions (if any), and the type of connection, are parameters which can take different values. Sets of values of these parameters are defined, both by the properties of the language, and by the method chosen for its description. In order to formulate each rule of the form just indicated, it is necessary to establish the values of its parameters. When separate rules are constructed, the order of their application should be specified. Thus, the analysis algorithm includes rules for finding the governing words, of the form described above, and rules concerning the order of their application.

The problem of the construction of an algorithm of syntactic textual analysis by a computer can be stated as follows:

> Given a text, the words of which possess certain characteristics; given also, the connections between the words and the types of these connections: construct an algorithm of analysis, such that, if the text in question were analysed by means of this algorithm, we could establish the same connections between words as those given.

The computer should, as it were, learn to analyse the text by the sample given.

In what follows, we shall say that the connections in the text are *established* for some of the words if their governing words are indicated.

The connections established in the text will be called *complete,* if to each word there corresponds a certain governing word. They will be called *regular* if conditions are satisfied such that, if we draw an arrow from the governing to the governed word, these arrows do not intersect nor form closed cycles.

As already mentioned, the connections between words are divided into types, and for the sake of simplicity we shall speak of "the type of connection of a word" when we mean the type of connection between this word and its governing word.

We shall define an *analyser* as a table of configurations in which each configuration consists of two words (the governed and the governing), an indication of their respective order in the text, and, an indication of the type of connection between them. We shall say that a pair of related words *realizes* a configuration, if this pair consists of the words indicated in the configuration, if their order in the text is as indicated, and if the connection between them is of the type indicated. We shall say that the pair of words of the text is *suitable* for the given configuration, if the words of the text coincide with the words of the configuration, if their respective order is that indicated but if the connection between these words is not ascertained.

The first step in the construction of the algorithm is to construct the analyser according to the given analysed text. After this, the computer will analyse the text with the help of this analyser, disregarding the connections established earlier, just as if no such connections were indicated at all in the text.

5. The Local Approach to Textual Analysis

Two principally different approaches to the analysis of a text are possible, and these may be tentatively called *local* and *integral*.

The first, or local approach is for the computer to to establish a certain variant of the connections between words of the text by analysis with the constructed analyser. It then compares the connections thus obtained with those indicated by the operator, marks the mistakes (the words where the computer got a wrong governing word or a wrong type of connection) and works out correction rules.

Several experiments with different languages have been carried out using this approach [6]). Different versions of the algorithm, altering the order of the steps, were studied. In the first of these algorithms, the words of any sentence were analysed from left to right, and for each word the governing one was first looked for to the left and then to the right. In the second case too, the sentences were analysed from left to right, but the governing word was looked for first to the right and then to the left, while in the third and fourth versions the sentences were analysed from right to left and the governing words were looked for as in the first and second versions respectively. Different versions gave approximately the same number of mistakes, but the second version proved to be somewhat better for the texts analysed.

So far the experiments have not been carried out with very long texts. One Russian, one English, one German and two French texts have been considered. These dealt with physics and mathematics. The number of mistakes with various versions of the algorithm are shown in Table 1.

Table 1

Text	Number of sentences	Number of words	Number of mistakes in each version			
			1	2	3	4
Russian	20	488	26	24	31	25
English	17	289	37	27	27	27
German	17	239	5	5	5	6
French (1)	20	444	31	28	30	29
French (2)	20	445	35	34	35	34

To types of additional rules were worked out in the experiments. The first kind of rule prohibits the establishment of a certain type of connection between two words, if the word which must be the governor in the pair is in a certain connection with its own governing word. Thus in the phrase “L’expression *F* est identique à la précédente”, the verb “est” was erroneously connected with the formula *F*. Generally speaking, such a connection is possible, but not in the case where the formula itself is related to a noun.

Examining the given analysed text, the computer decides what types of connection are possible if the word is a governing one in the realization of a given configuration.

The second type of rule prohibits the establishment of connections between words which are separated by certain “prohibitive” words. In accordance with the original text, the computer makes up a list of “prohibitive” words. In the experiments carried out, such elements as subordinate conjunctions, the semicolon, the parenthesis, etc., were designated “prohibitive”. The introduction of such additional rules leads to the correction of a number of mistakes, but the remainder require further additional rules.

As pointed out, in the experiments the words were substituted by sets of characteristics, these being taken as if morphological analysis had already been carried out. In addition, an analysis of homonyms was made for these texts. Some experiments were also carried out on the texts from which the homonyms had not been eliminated. Thus, in the second French text mentioned above, homonyms were analysed simultaneously with the establishment of connections, and the number of mistakes remained almost unchanged (there were 4 mistakes in the analysis of homonyms in the first version). So there are reasons to hope that the problems of analysis of homonyms can be solved along the same lines as those of syntactic analysis. Several experiments were carried out for developing some rules for the analysis of homonyms by the computer, rules which would operate independently of the analysis algorithm. These rules are described in the article by Korovina [7]).

6. The Integral Approach to Textual Analysis

The second possible approach to the analysis of the sentence, the integral one, consists in obtaining all admissible versions of analysis of the text, with subsequent selection. In accordance with the above, an admissible version of analysis is a complete regular set of connections, such that each connected pair of words is a realization of a certain configuration of the analyser. Algorithms, based upon such an approach have been developed by Iordanskaya in Moscow, and Zeitin and his group in Leningrad.

In the experiments carried out at the Mathematical Institute of the Academy of Sciences of the USSR, Vakulovskaya used a method of obtaining all admissible versions, the phrase being presented in the form of a matrix of 0s and 1s, as suggested by Sloutsker. Before explaining the construction of the matrix, it should be noted that we shall define as words *potentially governing* a given word, all those words of the same phrase which form, with the chosen word, a suitable pair for a certain configuration of the analyser.

Thus in the phrase “Ce groupe est alors un sous-groupe fermé du groupe linéaire d’un certain espace vectoriel”, potentially governing words for the word “fermé” are the words “groupe”, “est” and “sous-groupe”.

The original matrix for the phrase is constructed as follows. To each word there corresponds a row in the matrix. In this row, elements whose numbers correspond to the numbers of the words in the phrase which are potentially governing for the given word are taken to be 1, the rest of the elements in the row being 0. It may so happen that some rows of the matrix will contain several 1s, only one of which should be retained, i.e. one out of a number of potentially governing words must be chosen. The matrix for the phrase given above is shown in fig. 1. (Zeros are not indicated in the figures.)

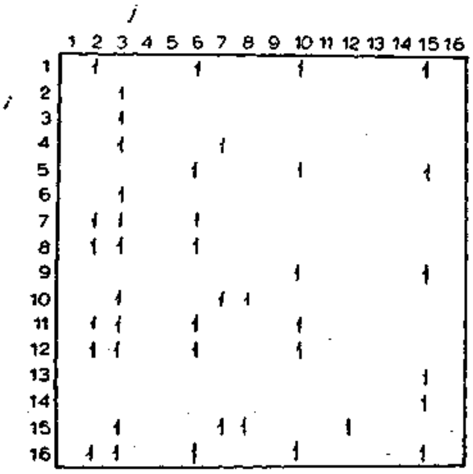

Fig. 1. Initial matrix.

The original matrix being given, it is necessary to construct a final matrix in which there is a 1 in each row to indicate the word which governs the given one. With such a representation of the phrase as a matrix, the conditions of regularity of connections can be reformulated as follows. Let us consider a row, where we can choose a 1 which is not located on the main diagonal. Let us assume that we have chosen a certain element, a_{ij}. We now construct a square, of which the vertex coincides with the chosen element, and the diagonal with the main diagonal of the matrix, and then we prolong the sides of this square (fig. 2). Then all 1s falling within the half-bands formed by the prolongation of the sides must be removed from the final matrix to which the chosen element a_{ij} belongs (they *contradict* a_{ij}). In fact, it is easy to see that these 1s correspond to connections which would intersect the connection between the *j*-th and *i*-th words. Besides, the final matrix must satisfy the condition of the absence of cycles. The conditions formulated naturally lead to the following method of construction of final matrices from the given initial one. It is evident that, if in a certain row of the initial matrix there is only one 1, then it must be present in any final matrix. First therefore, all the 1s which contradict those which are unique in their rows, are eliminated from the matrix; if new rows with unique 1s are formed, then any which contradict them are eliminated, etc. Then, if rows with several 1s are left, we take the first row and the first 1 in it, and eliminate all the elements which contradict if, then we take the second row, etc. As a result of this process we either obtain an empty row, i.e. no version of the analysis, or a final matrix which corresponds to a certain version of the analysis. The repetition of this process, successively treating all 1s in those rows which contain several 1s, gives all final matrices, i.e. all admissible versions of the analysis in the sense stated above. Now the task is to select a correct version. Evidently, if we only make use of the formal characteristics of the words, without reference to semantics, it will not always be possible to select a single version; but, in any case, the least possible number of versions should be left.

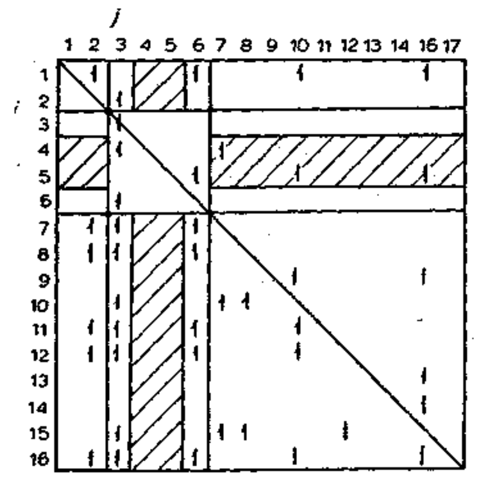

Fig. 2. Construction of final matrix.

7. Selection of Correct Versions

In the selection of versions both structural and statistical criteria are used. To statistical criteria there belong, for example, some functions of the frequencies of configurations taking part in a given version (the frequency of all configurations being indicated in the analyser).

The selection of versions according to structural criteria is carried out by examining the relations in the phrase, the order of examination being, in a sense, opposite to the order of construction. In the analysis of the phrase, the governing words are looked for, while in the process of selection and examination we find out which words turn out to be governed by the given word.

Thus, to structural criteria there belong:

1. Some words (as prepositions) must necessarily govern others, and therefore the versions of analysis in which a preposition has no governed word, are rejected.
2. When establishing the connection between the words of a certain pair, not only the characteristics of these words are essential, but also the type of connection of the governing word in the given pair. In other words, we verify the compatibility of the type of connection of the governing word with the types of connection of the words governed by it. This criterion corresponds to the first of the additional rules in the first local approach.
3. The compatibility of the pairs which turned out to be subordinate to one and the same word is verified. For instance, the French verb “est” can govern a noun (“est un sous-groupe” in the example given above) and it can also govern a past participle (“est fermé”) but not simultaneously.

Thus when constructing the original matrix, i.e. when finding out the words that potentially govern the given one, only the characteristics of two words are taken into consideration. When verifying the regularity of the version and the compatibility of the pairs or types of relation, or of different words governed by the same word, we consider not pairs but triplets of words.

In the example discussed above there were 44 versions. Among them 22 were such that the prepositions had no words to govern, and 17 contained incompatible pairs of subordinate words.

In the integral approach, concrete criteria of selection, as well as additional characteristics of configurations in the local approach are established with the help of the computer. In accordance with the given analysed text, tables are constructed in which it is indicated which words can have no subordinates, which words must necessarily have them, which types of governed word are possible with the given type of governing word, which pairs of types coexist with a certain type of governing word, etc. These tables are used in the selection of versions.

8. References

[1]) Yngve, V. H.: *A Progamming Language for Mechanical Translation.* Mechanical Translation, 5, I (1958).

[2]) Koulagina, O. S.: *Operator Description of Translation Algorithms and their Automatic Programming.* Problems of Cybernetics, 2 (1959).

[3]) Meltchouk, I. A.: *On the Standard Operators for the Algorithm of Automatic Analysis of Russian Scientific Texts.* Machine Translation, 2 (1961).

[4]) Meltchouk, I. A.: *Morphological Analysis in Machine Translation (on the Material of the Russian Language).* Problems of Cybernetics, 6 (1961).

[5]) Achmanova, O. S., I. A. Meltchouk, E. V. Padutcheva and R. M. Froumkina: On the Problem of Applying Precise Methods to the Study of Language. (Moscow 1961).

[6]) Koulagina, O. S.: *On the Use of Computers for the Construction of Algorithms of Text Analysis.* Problems of Cybernetics, 7 (1962).

[7]) Korovina, T. I.: *The Construction of Rules for the Elimination of Homonyms with the Help of a Computer.* Problems of Cybernetics, 7 (1962).