[*Mechanical Translation*, vol.5, no.3, December 1958; pp. 95-100]

The Work on Machine Translation in the Soviet Union *

Fourth International Congress of Slavicists Reports, Sept. 1958

V. Yu. Rozentsveig, First Moscow State Pedagogical Institute of Foreign Languages, Moscow, USSR

Problems of machine translation have been investigated in the Soviet Union since 1955.[1] A number of groups are carrying out theoretical and experimental work in the area of machine translation.

In the Institute of Precision Mechanics and Computer Technology of the Academy of Sciences of the USSR (ITM and VT) dictionaries and codes of rules (algorithms) have been compiled for machine translation from English, Chinese, and Japanese into Russian; and a German-Russian algorithm is being worked out. Experimental translations of individual passages have been made.[2] In the work of the ITM and VT group there is a marked striving for the rapid achievement of immediate, practical results. The efforts of this group are directed not so much toward a theoretical comprehension of the general problem of machine translation as toward a careful, detailed investigation of linguistic material, especially lexical. Dictionary routines, routines for analysis of the sentence in the source language, and routines for the synthesis of the sentence in the target language are being compiled in the ITM and VT on the basis of traditional methods of describing a language.

An essentially different course is being followed by the group working in the Steklov Mathematical Institute of the Academy of Sciences (MIAN). The problem of machine translation is being examined here as part of the larger problem of the automation of thought processes. The directors of this group regard the effective practical realization of machine translation only as the result of profound theoretical research in the area of mathematics and linguistics.

In MIAN three algorithms have been elaborated: French-Russian, English-Russian, and Hungarian-Russian.[3] During the compilation of the first of these algorithms in 1955-56, the workers in this group proceeded empirically, i. e. they extracted the rules for the translation of each word from a comparative analysis of French texts and their Russian translations. In the elaboration of the English-Russian algorithm, the MIAN group posed for themselves a more complex problem -- determination of the correspondences between the grammatical structures of two languages. The posing of such a problem was partially conditioned by the nature of the relationships of the English and Russian languages: although it was possible to build the analysis of a sentence on a morphological basis in translating a French mathematical text into Russian, such a method did not seem rational to the MIAN group in the case of English-Russian translations of similar texts. The problem was also partially conditioned by the theoretical goal of the director of the group. Professor A. A. Lyapunov: to work out strictly formal methods of describing languages in order to attain gradual automation of the whole process of machine translation.

---

* Translated by Lew R. Micklesen, Department of Far Eastern and Slavic Languages and Literature, University of Washington, December 1958.

---

1. The idea of machine translation was advanced even in the 30's by the inventor-technician, P. P. Smirnov-Troyansky.

---

2. I. K. Bel'skaya, "Concerning Certain General Problems of Machine Translation," Abstracts of the Conference on Machine Translation, Moscow, 1958, pp. 10-14, (hereafter referred to as Abstracts CMT).

---

3. See O. S. Kulagina and I. A. Mel'chuk, "Machine Translation from French to Russian, " Voprosy Yazykoznaniya, 1956, No. 5; T. N. Moloshnaya, "Some Problems of Syntax in Connection with Machine Translation from English to Russian, "Voprosy Yazykoznaniya, 1957, No. 4.

The theoretical basis for the isolation of typical sentence structures was the concept of the syntagma (according to de Saussure) or of the construct (according to Fortunatov). Machine translation, however, requires a certain modification of this system. In the structural syntactic analysis proposed by the author, T. N. Moloshnaya, of the English-Russian algorithm worked out at MIAN, constructs consisting not only of two members but also of many members (constructions with an absolute participle, etc.) are isolated. Such elementary structures were called configurations. They are composed of words classified according to formal signs. The analysis consists in reducing each configuration to its basic word, that is, shortening it. In this way, syntactical links are established between the words of a sentence. Synthesis of the Russian Sentence is made by means of substituting for it a given English configuration which corresponds to the Russian configuration and completing it with Russian words on the basis of the data of the dictionary, more precisely, of the Russian part of the dictionary, and on the basis of the corresponding morphological rules. The dictionary for machine translation, as compiled at MIAN during work on the French-Russian algorithm consists of two parts: (1) the foreign, containing the words of the given language (more precisely their stems, i.e. the graphically invariable parts of a word) with their corresponding tags indicating part of speech, idiomatic relationships, government by preposition and grammatical characteristics and (2), the Russian, containing Russian stems and the corresponding information about them. The Russian part of the dictionary is independent of the foreign part; so it may be used in translating from various languages. The rules for the morphological form of a Russian word are also independent of the language from which the translation is made.

The significance of the MIAN English-Russian algorithm lay in the fact that in contrast to all preceding algorithms in which the analysis of the text under translation was realized in terms of a translation into Russian (a category of the Russian language was ascribed to a foreign word), in T. N. Moloshnaya's algorithm the structural-grammatical analysis of an English sentence proceeded, in principal, independently of the language into which the text was being translated. This is extremely important, for an independent analysis opens the way for the realization of machine translation not only from one concrete language to another, but also from many languages to many others.

Several scientific groups are now working along this path opened up by the efforts of the MIAN Group. In the division of applied linguistics of the Institute of Linguistics of the USSR directed by A. A. Reformatsky, rules for the analysis and synthesis of a text and an abstract system of lexical and syntactic correspondences between various languages are being worked out independent of a translation into a concrete language by I. A. Mel'chuk. All of this should allow us to do machine translation from several languages into several other languages (the model of such an intermediary language is being made on the basis of an analysis of Russian, English, Chinese, French, and Hungarian). Syntactic analysis lies at the basis of the translation system being developed by I. A. Mel' chuk — morphological data are employed only as auxiliary data in the establishment of configurations, i.e. in bringing out the relationships between words in the source language and the expression of these relationships by means of the target language.

In this connection one should mention the research on the isolation and cataloguing of the system of relationships in the Russian language carried out in close collaboration with I. A. Mel'chuk in the Laboratory of Electrical Modelling of the Ail-Union Institute of Scientific and Technical Information of the State Scientific-Technical Committee in the Soviet of Ministers of the USSR and of the Academy of Sciences of the USSR(LE). In Russian mathematical texts the workers of this laboratory, Z.M. Volotskaya, E. V. Paducheva, I. N. Shelimova, and A. L. Shumilina isolated and described about 200 syntagmas (two-membered constructs in a subordinate relationship) which are essential in both the analysis and the synthesis of a Russian sentence.

A substantial contribution to the theory of translation algorithms and their programming was made by O.S. Kulagina (MIAN). She developed a system of so-called elementary operators of the simplest steps of which any translation process may consist and of programs corresponding to these steps. As a result, significant generalization and standardization in the process of making algorithms can be attained, all of which allows us to pose the problem of automation of the programming of algorithms and then the problem of their automation and construction.

The Experimental Laboratory of Machine Translation of the Leningrad State University (ELMP) under the directorship of N.D.Andreyev is also endeavoring to realize the idea of developing completely independent methods of analysis and synthesis and of some abstract logical system making it possible to go from analysis to synthesis, i.e. a system that will serve as an intermediary language. In this laboratory extensive material from various linguistic systems is being investigated; Indonesian-Russian, Arabic-Russian, Japanese-Russian, Burmese-Russian, Norwegian-Russian, English-Russian, Spanish-Russian and Turkish-Russian algorithms are being developed. The intermediary language which N. D. Andreyev is attempting to create is an artificial language constructed by averaging the phenomena of various languages. It is regarded as a material language with its lexicon, its morphology, and its syntax, but with the one peculiarity that it consists of symbols *. In the selection of the categories at the basis of his symbolization, N. D. Andreyev considers the most frequent phenomena and also the international prestige of each language.[4]

The system of signs developed in ELMP for the recording of the intermediary language can be used also for the recording of information in information machines.

Along with work on the algorithms of machine translation from foreign languages into Russian and from Russian into foreign languages being conducted in the Gorki State University, the following algorithms are being elaborated: Armenian-Russian and Russian-Armenian (in the Computation Center of the Academy of Sciences of the Armenian SSR), Georgian-Russian and Russian-Georgian (in the Institute of Automation and Telemechanics of the Academy of Sciences of the Georgian SSR).

In the First Moscow State Institute of Foreign Languages (I MGPIIYa) where under the directorship of I.I.Revzin theoretical investigations of the problems of machine translation and of related problems of linguistic theory of translation and methodology of foreign language teaching have been carried out, the elaboration of Russian-English, Russian-French, and Russian-Spanish translation algorithms for foreign policy texts has begun . At the Institute, the Machine Translation Society has been created at whose meetings theoretical problems are discussed and an exchange of ideas about the practical problems of the compilation of the algorithms takes place. In the bulletin published by the Society are published both theoretical and experimental work connected with the problem of machine translation. In May, 1958, the Society convened the First All-Union Conference on Machine Translation. Seventy-nine institutions were represented at the conference, including twenty-one institutes of the Academy of Sciences of the USSR and eight institutes of the Academies of Science of the Union Republics, eleven universities, and nineteen other institutions of higher learning in the country. Linguists, mathematicians, and technicians took part in the work of the conference. At the plenary and sectional meetings of the conference there were discussions of more than seventy reports and communications devoted to general linguistic problems arising in connection with the use of language in present-day automatic devices as well as to special problems of construction of algorithms for machine translation. [5]

The central problem now confronting linguists working in the field of machine translation is that of the methods of formal description of linguistic structures. Structural methods, particularly the methods elaborated by descriptive linguistics, offer much of value for the formal description of language — it was not by accident that the work of Fries in the structure of the English language proved useful in working out English configurations. It has become clear, however, that these methods are inadequate for the formal description of language to the extent that this is demanded in automatic translation. In connection with this a search for means of applying mathematical methods to the analysis of language was begun. With this in mind the Department of Philology of the Moscow State University initiated a seminar on mathematical linguistics in 1956, joining mathematicians and linguists under the direction of P.S. Kuznetsov, V. V. Ivanov, and V. A.Uspensky. Here, as well as at the meetings of the Machine Translation Society the idea, suggested by Academicians A. N. Kolmogorov and A. A. Lyapunov, of applying the methods of mathematical logic and of set

---

* Translator's note: The author obviously means symbols different from the conventional symbols of language.

4. N. D. Andreyev, "Machine Translation and the Problem of an Intermediary Language," Voprosy Yazykoznaniya, 1957, No. 5.

5. See Abstracts CMT, M., 1958

theory to the study of language was discussed. Thus, for example, A. N. Kolmogorov's idea about the possibility of a strict formal definition of the category of case (the work of V.A. Uspensky and, in part, also of R. L. Dobrushin) was expounded and developed. It is interesting to note that eight cases can be counted in the declensional system of the Russian substantive according to this definition.

A method for defining grammatical categories, worked out by a student of Professor Lyapunov, O. S. Kulagina (MIAN), was discussed at the seminar. This method of definition allows one to obtain, independently of the concrete features of the language, a classification of words and a determination of their syntactic relationships. Language in this conception is regarded as a set of elements — words, or more exactly — word forms. A finite number of words arranged in a definite order is called a sentence. Certain sentences are assumed to be marked — these are sentences constructed according to the norms of the given language — others are unmarked. According to the criteria of mutual substitutability of words in the marked sentences the entire set of words is broken down into groups of mutually equivalent words.

In terms of this system a series of definitions corresponding, in general, to certain traditional morphological categories, for example, parts of speech, was successfully obtained. The advantage of this classification lies, however, in the fact that it has been deduced on the basis of an exact and strictly formal system of definitions. It is particularly effective for languages with a rather symmetrical system of word forms (for example, French). In languages like Russian that do not possess this symmetry, the method of defining a grammatical category proposed by R. L. Dobrushin can be utilized.

By making use of the criterion of equivalency, the relationships between the classes of words isolated are also determined. Moreover, the concept of configuration, mentioned earlier, gets a more exact definition: a configuration is defined by O. S. Kulagina as that combination of not less than two words belonging to various non-intersecting subsets, which can be reduced to one element without any marked sentence containing this configuration losing its marked quality. Thus the combination of the words "thick book" in the sentence "the thick book lies on the table" can be reduced to the element "book" or can be replaced by the element "thing" or the element "it" without the sentence ceasing to be marked. The isolation of the configurations allows one to determine the syntactic structure of the sentence.

The set-theory concept of language is strictly deductive and formal. This is just what determines its importance both for general linguistics and for machine translation. Naturally the formal description of language is possible only to a limited extent. Thus, the concept of the marked quality of sentences, without which it is impossible to determine the equivalence of elements and configurations of a language, will have little effect if it is extended to all functional areas of language. But in a limited sphere of language — and machine translation at the present time is being considered only within the limits of scientific and technical prose — this concept is sufficiently exact and effective. Thus, all sentences in a given language which are met in a given field of scientific literature can be considered marked.

The set-theory conception of language is important in yet another respect. Since it allows us to construct and investigate a grammatical model, i.e. a simplified analog of actual linguistic relationships, this theory opens one of the possible ways for logico-semantic investigations of language. In this connection we should point to the ideas of V. V. Ivanov about the possibility of applying mathematical methods to the definition of the lexical meaning of words. I note that, contrary to wide-spread opinion, the theory of machine translation is not limited to the investigation of language in its formal aspect alone. The search for methods of objective, precise description of the system of meanings in language has begun.

If it is true that complete formal description of an actual language is hardly accessible, that it is necessary to attain only formal approximations to actual language, then a statistical evaluation of the probability of this approximation acquires special importance [6]. On the other hand, certain phenomena of language do not yield, for the time being, to structural description and can be formally described only statistically.

---

6. See V. A. Uspensky, "Conference on the Statistics of Speech," Voprosy Yazykoznaniya, 1958, No. 1, p. 173.

The quantitative aspect of linguistic phenomena, both lexical and grammatical, has been considered, as a rule, in all the algorithms formulated. One should point particularly to the statistical investigations carried out on Russian language material in the Laboratory of Electrical Modeling. I have already mentioned the cataloguing of Russian syntagmas. This work was accompanied by a statistical investigation of the language of Russian mathematical texts. The results of this work conducted by I. A. Mel'chuk, T. N. Moloshnaya, A. L. Shumilina, Z. M. Volotskaya, and I. I. Shelimova, were, along with other works, announced at the conference on the statistics of speech convoked in October 1957 by the Section of Speech of the Commission on Acoustics of the Academy of Sciences of the USSR and by Leningrad University. This work is of interest not only in a practical respect. Its value consists in a true solution to the problem of combining statistical and structural methods: a count of linguistic elements was carried out by the authors on the basis of a clear-cut definition of such concepts as "syntagma", "type of syntagma", etc. As I. I. Revzin showed in his report presented at the conference mentioned, the correlation of structural and statistical methods has a two-sided nature: statistics aids in specifying the structure of language and an exact structural definition of units, the number of which are counted, insures the proper conduct of the statistical investigation.

A frequency count of dictionary units is important not only in connection with machine translation. No longer speaking about statistical investigations of problems of general and particular linguistics [7], which have already become traditional, we shall point to recent works connected with the use of language in various devices for the storage, processing, and transmission of information. In reference to the Russian material we can call attention to the use of methods of machine translation for the coding of telegraphic and telephonic messages.

It has been established (V. I. Grigor'ev and G. G. Belonogov) that the size of a telegraph message in Russian can be diminished by 3-4 times if the telegraphic communication is translated from a letter code into a dictionary (lexical) code. Statistical investigations have shown that in the case of such coding 4, 000 common words would be sufficient in order to insure the transmission of 97.5 percent of a general-language text.

The problem examined here is connected, for the most part, with an analysis of the text under translation. For the Soviet specialists the elaboration of effective methods for analysis presented special difficulties: they dealt primarily with morphologically poor languages. It would be erroneous, however, to assume that the synthesis of the Russian sentence did not present any serious difficulties to them. By way of illustration we may cite the difficulties arising in the synthesis of Russian aspectual forms, inasmuch as the category of aspect permeates the entire Russian verbal system.

Here two problems of principle arise. In the first place, it is necessary to find a principle of classification of Russian verbs which will allow us to obtain for each verb in an absolutely regular way (by adding or taking away the same letters) all forms of the perfective as well as of the imperfective aspect. Such work was done by Z. M. Volotskaya (LE), who obtained three breakdowns of the whole Russian verbal complex according to method of formation: a) of present tense forms; b) of past tense forms; and c) of the perfective stem from the imperfect stem. [8]

In the second place — and this task is much more difficult — it is necessary to work out the rules for the choice of one or the other aspectual form. Inasmuch as the tendency towards carrying out the operations of synthesis independently from those of analysis has already been noted, these rules must be constructed on the basis of contextual data, considering, for example, the presence in the sentence of adverbs, the character of the combination, etc. In a series of cases one must limit oneself only to a probable solution, based on statistics.

The problem of machine translation from Russian, of course, occupies Soviet investigators less than the problem of translation into Russian. But investigative work connected with the analysis of the Russian sentence has already begun (chiefly in the Laboratory of Electrical Modeling, the Division of Applied Linguistics of the Institute of Linguistics of the Academy of Science of the USSR and in ITM and VT). From the point of view of general linguistics the work reveal-

7. In this connection one should recall the works in the statistical investigation of Russian literary works, carried out in the 20's and 30's by A. I. Peshkovsky, M. Peterson, et al.

8. See Abstracts CMT, p. 87

ing the redundancy of certain categories of the Russian language is most interesting. Thus, for example, the category of gender in the Russian verb, expressed only in the forms in -1 of the singular of the past tense and of the conditional mood, is redundant, unnecessary from the standpoint of analysis. It is clear (V. N. Vinogradova, the Institute of Linguistics of the Academy of Science of the USSR) that in scientific texts the number of verbs with the expressed form of gender comprises from four to thirty percent and that in the majority of sentences the verb can be related only to the subject — the only substantive in the nominative case. Nor is it necessary, in most cases, to consider the inflection of the Russian adjective and determine the relationships of the adjective to the substantive with which it agrees on the basis of the position of the adjective in the sentence. (N. N. Leont'eva and G. H. Vavilova, the Institute of Linguistics).

Interesting also is the work on the determination of syntactic links for the preposition-case groups of the Russian language (I. N. Shelimova) and also the work on the elaboration of the syntactic links for formulas in Russian mathematical texts (M. M. Langleben) — by formulas the author means all elements not found in the machine dictionary during the processing of the text (mathematical formulas, foreign-language citations, surnames, etc.)

For the analysis of a Russian sentence it is necessary to characterize the marks of punctuation. Only in such a way can one find the limits of a simple clause within a sentence, isolate its similar members, aid the further clarification of the co-relationships of the individual parts of a sentence with complex punctuation, determine a group of similar members. T. N. Nikolayeva (ITM and VT) conducted an analysis of polysemantic marks of punctuation (comma, dash, colon) in Russian[9].

Thus the realization of machine translation presupposes serious theoretical investigations, which, in turn enrich the problems of general and applied linguistics.

---

9. See Abstracts CMT, pp. 104-107