

SOME SOVIET PAPERS ON MACHINE TRANSLATION¹

1. D. Ju. Panov, "Avtomaticeskij perevod"*, *Moscow Popular Scientific Series* (Moscow, Acad. of Sc. USSR, 1956), pp. 5-45.
2. P. S. Kuznecov, A. A. Ljapunov and A. A. Reformatskij, "Fundamental Problems of Machine Translation"*, *Voprosy jazykoznanija* 5.107-11 (1956).
3. O. S. Kulagina and I. A. Mel'čuk, "Mašinnyj perevod s francuzskogo jazyka na russkij", *Voprosy jazykoznanija* 5.111-21 (1956).
4. I. S. Muxin, *An Experiment of the Machine Translation of Languages carried out on the BESM* (Moscow, Acad. of Sc. USSR, 1956). 28 pp.
5. D. Panov, *Concerning the Problem of Machine Translation of Languages* (Moscow, Acad. of Sc. USSR, 1956), 36 pp.
6. T. N. Mološnaja, "Nekotorye voprosy sintaksisa v svjazi s mašinnyj perevodom s anglijskogo jazyka na russkij", *Voprosy jazykoznanija* 6.92-7 (1957).
7. I. K. Bel'skaja, "Machine Translation of Languages", *Research* 10.383-9(1957).
8. D. Ju. Panov, A. A. Ljapunov and I. S. Muxin, "Avtomatizacija perevoda s odnogo jazyka na drugoj", *Sessija Akademii Nauk SSSR po naučnym voprosam avtomatizacii proizvodstva, 15-20 oktjabrja 1956 g., Plenarnye zasedanija* (Moscow, Acad. Sc. USSR, 1957), pp. 181-214.

Machine translation in the Soviet Union is now definitely in the operational stage. The material before me clearly indicates that theoretical discussion was cut short rather early in the game and a decision was made in favor of a structuralist approach to the linguistic phase of the problem: "Since machine translation emphasizes the communicative function of language and since the machine must determine linguistic relationships in a purely formal sense, there is some justification for adopting a linguistic approach that is often condemned as 'formalism' or 'structuralism'... The work of formalizing relationships between languages must rely on those linguistic theories that are based on the formal aspect of language study. In modern linguistics these theories are represented by the various schools of structuralism." (2.)

The research design chosen is one of pragmatic experimentation, with definite and

¹ The papers marked with an asterisk were available in English translation only; in quoting from them, no page references can be given and only the serial number is indicated, thus: (1.) Quotations from the remaining papers will be given by serial number and page reference, thus: (3.113) = third item (O. S. Kulagina and I. A. Mel'čuk), p. 113.

limited starting points: "In view of differences among languages there can be no universal system of translation rules; each system of rules can relate only 2 languages in one direction." (2.)

The technical discussion that I have been able to follow most clearly concerns French into Russian (3., 8.182-7) and English into Russian (4., 5., 6., 8.187-207); German into Russian (8.208), Chinese (7.387, 8.210) and Japanese (7.387, 8.211) have also been mentioned.

Before commenting on it, I should like to summarize briefly some of the crucial problems of machine translation.²

The purpose of machine translation research is to develop a computer program capable of effecting translation decisions automatically.

Such a translation program can be divided into two major phases: a data-processing phase and an internal phase.³

The data-processing phase consists first of a sensing-and-matching routine by which input items are sensed (= read) and matched against the entries in a previously prepared and stored glossary. This is the dictionary lookup. Secondly, the data-processing phase must establish a link between the dictionary lookup and the internal phase in which the real work of decision-making is done. This link is usually accomplished by storing an appropriate code under the entry words of the dictionary, which is utilized in the internal phase to call for appropriate routines and allow their execution.

The internal phase is concerned with the implementation of the actual translation decisions. These decisions are either selection decisions - concerning the choice of one out of several possible translation equivalents, or arrangement decisions - concerning the rearrangement of the translations of individual input words, necessary to meet the word order requirements of the output text.

In order to solve a given decision problem, the first step is to ascertain the decision point - that is, the input item best suited as a starting point for the routine designed to solve the problem. The second step is to search the context of the decision point for decision cues - those elements of the input text which yield the information needed to make the particular decision. These two steps together constitute what I call the recognition routine; its purpose is the recognition and utilization for decision-making of the linguistic function of formal graphic elements. I use the term "search span" to cover the extent of text covered by the recognition routines of a particular program.

A third step is required: once the translation decision is made, the necessary instructions are generated to produce an output based on this decision. This constitutes what I call the command routine.

The several Soviet groups (two are mentioned in 8.182, additional groups are said to exist) seem to make a clear-cut separation between the recognition and command routines, which they call "source-language analysis" and "Russian synthesis" respectively (3.117-21, 4.11, 5.10-11, 7.389, 8.185, 8.196).

The recognition routines which I have seen seem to involve short search spans. Let me cite as an example Panov's routine for "look" (5.20, simplified by me):

(a) check whether the following word or next following preposition is "for"

² For a more detailed discussion, see my "Some Linguistic Problems in Machine Translation", *For Roman Jakobson* (The Hague, 1956), pp. 180-6, and "Machine Translation", *Reports to the VIIIth International Congress of Linguists, Oslo, 1957*, vol. I, pp. 103-11.

³ For advice in formulating the statements relating to computer operations and programming, I am indebted to Dr. R. J. Arms of the Division for Applied Mathematics, National Bureau of Standards, Washington, D. C. I am, of course, solely responsible for my opinions and conclusions.

yes - take step (b)

no - take step (c)

(b) go into appropriate command routine for translation by proper form of *искать*

(c) check whether the immediately following word is "upon"

yes - take step (d)

no - take step (e)

(d) go into command routine for translation by proper form of "смотреть"

(e) go into command routine for translation by proper form of "рассматривать" and insert "как" immediately after the form chosen.

For translation decisions affecting multiword sequences, the decision point is determined by ascertaining the "major component" of a "grammatical configuration" (8.188-9). This is defined as the "pivotal, grammatically dominant word to which the remaining words are subordinated"; it allows "the consistent determination of the relations between all the words of the sentence" (8.189).

The command routines seem to concern primarily the selection of appropriate stem and suffix allomorphs for making up the Russian output.

As an example, Kulagina and Mel'čuk's recognition routine yields "рассматрива-" and "рассмотр-" as the equivalents of French "considerer", with command # 5 for stem alternant choice. This command requires the selection of "рассматрива-" for imperfective, "рассмотр-" for perfective forms. Similar suffix alternant choice commands are given on the basis of information yielded by the recognition routine as well as of the requirements of the stem alternant choice.

The two types of programming techniques employable in machine translation are table-lookup and algorithmic.⁴ A table-lookup operation is in essence one where the answer to a question is looked up in a table stored in memory; an algorithmic operation is one where the answer is computed by a succession of simple arithmetic operations based on a series of yes/no questions in the shape of a logical tree.

The nature of the data-processing phase of machine translation is such that only a table lookup can solve the basic problem of dictionary searching. The internal phase, on the other hand, in theory admits of both table-lookup and logical-tree operations, and different approaches to machine translation vary in terms of their employment of the two techniques.

Our Soviet colleagues seem to favor the logical tree for both aspects of the internal phase (for the recognition routine, see the example from Panov quoted above; for the command routine see especially 8.204-6).

Styles of formulation vary; some of the sample flow charts I have seen are in block-diagram style (especially 1.); other samples (they seem to be later ones) show step-sequence style (5., 7., 8. *passim*).

The major difference between the work of the Soviets and American approaches (discounting divergences of style and other matters of preference) lies, in my opinion, in the difference of objective. They are translating *into* Russian, we are translating *from* Russian. This means that in essence the problems which they face in their recognition routines, we face in our command routines, and conversely.

I am thinking specifically of the translation problems arising from the formal marking of the major clause functions: those of subject, predicate, object. In Russian, these clause functions are marked primarily by paradigmatic morphs, the morphemic function of which must often be ascertained from the context, and only in cases of paradigmatic ambiguity will conditions of word order and special lexical conditions mark the

⁴ Gilbert W. King, "The Requirements of Lexical Storage", *Georgetown University Monograph Series on Languages and Linguistics*, No. 10, p. 79 (Washington, D. C., 1957).

clause function. In English and French, on the other hand, the clause functions are primarily indicated by order.

A recognition routine for Russian clause functions must therefore search the span of the entire clause for appropriate morphs and check out their functions. The corresponding command routine for either English or French must be capable of accomplishing the rearrangement of clause-member-sized blocks of words over non-trivial distances.

On the other hand, a recognition routine for English or French clause members may proceed linearly to identify clause functions by order of appearance in the input. The corresponding command routine for Russian must then generate different paradigmatic morphs (and stem alternants when necessary) for formally same French and English blocks of words in different locations within the clause.

Thus, a translation program for Russian into French or English requires both significantly greater search spans and significantly more extensive and complex rearrangement commands than a program in the opposite direction. In my own work,⁵ I have found that the extension of the search span and the proper implementation of large-scale rearrangement constitute the major, though not insurmountable, difficulties.

For the general linguist, it will be interesting to note in this context that the old-fashioned typological categories of synthetic versus analytic have acquired a certain operational validity in machine translation research.

Institute of Languages and Linguistics
Georgetown University

Paul L. Garvin