MACHINE TRANSLATION IN THE SOVIET UNION

I. A. Melchuk

Moscow, U.S.S.R.

(Based on a translation from Vestnik Akademii Nauk U.S.S.R., No. 2, 1959)

Work on machine translation began in our country in 1955, at the Precision Mechanics and Computational Techniques Institute and the V.A. Steklov Mathematical Institute of the USSR Academy of Sciences. They were later joined by the Linguistics Institute of the USSR Academy of Sciences; the Leningrad University, which set up a machine-translation laboratory; the Computing Center of the Academy of Sciences of the Armenian Soviet Socialist Republic: the Electronics, Automation and Telemechanics Institute of the Academy of Sciences of the Georgian Soviet Socialist Republic; and other institutions.

Compilation of "Algorithms"

At the first stage, the work consisted mainly of compiling the so-called algorithms of machine translation. For a machine to translate a text from one language into mother all the necessary operations must be given as a special set of rules. These rules must be precisely formulated and perfectly clear; must permit of mechanical performance; and constitute a logically coherent system providing for all possible cases.

A number of algorithms were compiled in 1955-1957: a French-Russian, two English-Russian, a Chinese-Russian, a German-Russian, a Japanese-Russian, a Hungarian-Russian. The French-Russian and one English-Russian were programmed and tested on computers; that is, translations were made of passages from French and English scientific (mathematical) texts into Russian. The other algorithms are in the programming stage.

At present, work on machine translation is proceeding along the following three main lines: investigation of possible ways of machine translation for selection of the best; development, by close collaboration of mathematicians and linguists, of precise (primarily mathematical) methods of language description; and study of interconnection between machine translation and other practical applications of linguistics with a view to generalizing and most fully applying the results achieved in allied fields.

Problem: To Work Out General Theory

What are the basic problems of this work?

Whereas three years ago the cardinal task was that of making up algorithms and applying them, now that we already have composed and applied algorithms, chief attention is focused on generalizing the results obtained.

The primary task now is to work out a general theory of compiling translation algorithms (notably, to work out a universal form of translation algorithm as such, and rules of applying the form to concrete languages). When this task is successfully accomplished, it will become possible to have the machine itself compile translation algorithms on the basis of glossaries and parallel texts prepared in advance.

A System of Operators

Of great importance for the general theory of translation algorithms is a system of operators worked out at the Mathematical Institute. Under this system the translation process is broken up into a number of standard acts which cake place in translating from any language into another. Such an elementary act, together with its corresponding standard computer program, is called an operator. Any algorithm may be represented as a sequence of operators. Recording algorithms in terms of operators makes it possible to mechanize the highly laborious process of programming translation algorithms. Thus, in programming a part of the Hungarian-Russian algorithm, five programs were compiled within five minutes, which, if done in the usual way, would require from 20 to 30 man-days. Operator recording is also very important theoretically since it makes it easier to unify algorithms and work out a single universal form of translation algorithm.

Next, An Electronic Editor

Today machine translation is regarded only as the first stage toward solving a more general and more important problem: by most fully using electronic machines as auxiliary tools of human thinking, to make the machine capable of performing the widest possible operations with texts written in different languages, to enable it not only to translate but also to edit, make abstracts, furnish bibliographical and other references, etc. All these operations boil down to extracting from the text required information and to recording that information in some other form. To carry out these operations a special "language" is needed in which the information from the text would be recorded.

Such a language should (1) ensure a simple and accurate recording of the extracted information, and (2) be convenient for translating into it texts written in natural languages. This language can be used both for recording and storing information in the machine (the language in which the information is recorded for stor-

ing in the machine is called "information language"), and as an intermediary for machine translation. In translating from many languages into many others in any direction it is possible to translate from the source language into the intermediary language, and from the latter into the target language. This makes it possible not only to reduce the number of algorithms necessary for direct language-to-language translation, but also makes it easier to unify them. The basic problem of most work on machine translation in our country now is to build up an intermediary language.

Of the many possible ways of building up an intermediary language, two are being actively investigated, and it is one them that we shall dwell.

A New Language?

One consists in producing an intermediary language as some artificial language possessing its own vocabulary, morphology and syntax (i.e., similar, to natural languages or artificial languages like Esperanto). The components of the intermediary language are determined by statistical investigation of the languages in question: only those phenomena are imparted to the intermediary language which are widespread in all or most of these concrete languages, with each allowed a share in proportion to the number of people speaking it.

Such an intermediary language will be an "intersection" product, as it were, of a number of given (natural) languages, for which a system of symbols has been worked out. In the future this formal-logic system is expected to be used as an information language.

The other way is to construct the intermediary language only as a system of correspondences between natural languages. The correspondences are established at three levels: vocabulary (between words and idioms of various languages), morphological and word-building, and syntactical (between elementary syntactical constructions).

Translation equivalent words of different languages (bundles of lexical correspondences) form sets and these sets constitute the words of the intermediary language; its syntactical relations are bundles of syntactical correspondences. The whole thing boils down to the following: It is assumed that the intermediary language is an "aggregate" of all the languages under review; this means that any differences occurring in all these languages may be expressed in the intermediary language. May, but not must; they are expressed not obligatorily, but on occasion, if they occur in the source language.

Bundles of morphological correspondences are regarded as words of the intermediary language (like the bundles of vocabulary correspondencies). These words may and may not occur: thus, the noun number category will be expressed in the intermediary language when translating from languages where it exists (Russian, English, Armenian, Hungarian, etc.), and will not be expressed when translating from Chinese where this category does not exist. This has been done in order to avoid losses of relevant information and to avoid producing superfluous information, no matter what pair of languages are involved in translation.

The intermediary language obligatorily expresses only two kinds of differences: lexical and syntactical (words and relations between them), i.e., differences which exist in all human languages and without which any language is unthinkable.

On the whole, the proposed intermediary language to certain extent resembles, on the one hand, the so-called Uhrsprachen of comparative linguistics (which likewise constitutes a system of correspondences between languages), and, on the other hand, calculuses of mathematical logic ("words" and "syntactical relations" of the intermediary language correspond to the alphabet and the formation rules of formal-logic languages).

Translation by Tables

The intermediary language as a system of correspondences can be set up in tables with vertical columns and horizontal lines. The columns are assigned to different languages; each line, to translation equivalent units of different languages. The numbers of the lines containing lexical and morphological equivalents represent words of the intermediary language; the numbers of the lines containing syntactical correspondences represent its syntactical relations.

The process of translation by means of an intermediary language is divided into two phases: analysis, or translation from the source language into the intermediary language, i.e., the numbers of the respective lines in rhe tables, are, by means of special routines, referred to various units of the source language; and synthesis, or translation from the intermediary language into the target language, i.e., units of the target language, selected from bundles of correspondences, are given the proper morphological forms and lined up in accordance with the laws of the target language.

A model of an intermediary language is now being developed for short passages from mathematical texts. Algorithms of independent analysis and synthesis are being devised for a number of languages; work has begun on establishing word correspondences between the major European languages.

Syntactical analysis, as a result of which syntactical connections between all the words of the translation text are determined, is the central part of an algorithm for machine translation. This is done by means of a list of elementary syntactical constructions (configurations) occurring in the texts of the given type, and by means of rules for detecting them in the text. Therefore, to build up an algorithm it is necessary to have sufficiently full lists of configurations for all the languages used.

Language Peculiarities Cause Problems

There are a number of other, purely linguistic, problems, the solution of which is necessary to construct an algorithm and which requires independent research. The

latter includes, among others, the problem of finding redundancies in a language, i.e., historically-evolved categories which in the system of a modern language perform no meaningful function. Thus, the gender category of the Russian verb has become almost entirely redundant, the inflections of Russian and French adjectives are largely redundant, the form differences of the Russian dative and local cases are always redundant and of the nominative and accusative cases are nearly always redundant, etc. The problem of redundance in a language is of great importance also for communications engineering, as the elimination of textual redundances makes it possible to increase many times the effectiveness of transmitting and receiving devices. Therefore, the efforts of machine translation specialists and communications engineers are being pooled to solve this problem.

Statistical Approach Necessary

Linguistic research on machine translation must be based on many-sided statistical investigation of the text. Statistics are necessary to limit the material under investigation, to isolate the range of phenomena to be described and systematized. Quantitative characteristics make it possible to appreciate the specific weight of various language phenomena in order to concentrate attention on essentials, leaving aside secondary aspects; they are also needed to assess the efficacy of one or another solution. Lastly, since absolutely precise solutions of one or another linguistic problem are not always possible, statistics help to find approximate, more plausible solutions,

Statistical description of speech is of considerable interest not only for machine translation, but also for communication engineering, printing, language teaching methods, etc. It is therefore a primary task to carry on statistical investigations in different languages, Russian first and foremost, on an appropriately large scale.

For these investigations to be effective it is necessary to use widely analytical and electronic computing machines, which again calls for close contact of linguists with specialists in other respective fields.

Specialized "Language" Computer

In conclusion, we should like to mention one more field in which linguists, mathematicians and electronic engineers should cooperate: designing of special translation and information machines for all kinds of work connected with language. (Up to this date in our country and abroad experimental translations are made on general-purpose computers not adapted for this purpose.)

Coordination of research in all these lines was in no small measure facilitated by the First All-Union Conference on Machine Translation, held in Moscow in May 1958.

Imparting Human Speech to Electronic Machines

All research on machine translation should be regarded as the initial stage of a wider range of work the goal of which is to impart human speech to electronic machines. Achievement of that goal will produce a real revolution in science and technology. And solution of the machine translation problem now directly confronting researchers will be a step forward toward that goal.