TRANSLATING RUSSIAN BY MACHINE

KENNETH E. HARPER

The idea of using machines to translate foreign languages has been given serious consideration only within the past few years. This new area of research has opened up with the development of modern high-speed electronic computers. These machines can solve in minutes or hours complex mathematical problems which humans require years to solve. Since mathematics is itself a language—a set of symbols used to communicate thought — why can't computers be used to translate French into English, or Chinese into Portuguese? In order to do this, we must discover what the rules of the language are which make meaningful sentences from a collection of words, and build a machine to use the rules to translate the words.

Can languages be resolved into a set of rules for mechanical processing? The basic assumption of mechanical translation is that this can be done. A written English word, for example, is a group of alphabetic symbols to which a certain meaning has been given by English-speaking people; sentences are composed of these words and are given meaning both by the arrangement and form of the words. Mechanical translation (or MT as it is familiarly called) assumes that these characteristics of words (meaning, arrangement, and forms) can be subjected to routine analysis. Thus, a given word has a predictable meaning, either by itself or in conjunction with surrounding words; if its position in a sentence affects the meaning of the sentence, there exist some "rules" which prevent ambiguity. "I gave the bone a dog" is ambiguous, or meaningless, in English because the rules of word order have been violated; on the other hand, this order is possible in Russian, because the rules of noun declension prevent ambiguity. Language cannot function efficiently without rules; these of course differ widely in different languages.

Kenneth E. Harper—Mr. Harper, of the University of California at Los Angeles, began his Russian studies during World War II as a language officer in Naval Intelligence. The device he describes here was first worked out by Mr. Huskey of the Institute of Numerical Analysis who designed SWAC, the automatic digital computer, and conceived of adapting it for that purpose. Mr. Harper's article is reprinted from *Idea and Experiment, A University at Work, 3, 3* (March, 1954) 10-13, a periodical edited and published by the Faculty of the University of California.

If rules exist for human use, we may be able to define and utilize them mechanically. At first glance, this notion appears ridiculous, or even pernicious; It seems to ignore the human factor, and to place the cathode ray tube on a par with the brain cell. Obviously, there are limitations to what can be done; at the same time, it is entirely possible that written language is more rigidly circumscribed by rules than we like to believe. Research in MT has already discovered many routines in sentence structure of Russian and German which have merely been taken for granted. In the process of translation, we shall have to take into account the routines observed in the original language, and adapt them to the different routines which may be observed in the "target language."

It is clear that mechanical gadgets cannot cope with the "subjective" quality of written language. Humans are not always efficient in their attempts at communication (people *do* misunderstand each other); in addition, language is not always used for the direct communication of thought (literature thrives on misdirection and suggestion). However, writing which has the single purpose of conveying information tends to be more direct. It is less subject to the stylistic deviations of the writer. By and large, this is true of scientific writing. A scientific book or article contains a relatively simple vocabulary and sentence structure, and is thus ideal for experimentation in mechanical translation. The first experiments in MT are therefore being made with scientific literature. (Translations from Russian to English are the first order of business, because of the large amount of scientific literature in Russian generally unavailable to American scientists.)

The actual procedure of MT will vary according to the system used. For purposes of translation, the computing device may be compared to a college freshman studying Russian. Tests have shown that a student who is familiar with the Russian alphabet can translate scientific Russian adequately if equipped with a dictionary and a set of grammatical rules. In order to do this, the student must follow precisely the instructions set down for him. Naturally, he works very slowly. The computing machine is like the student: it will do anything it is told to do, following the prescribed procedure, and will come up with the same result. Its advantages are accuracy and speed, for it will perform operations in micro-seconds, rather than in minutes.

grammatical analysis, and output. The Russian sentence is fed into the machine word by word, by means of various devices — a typewriter, Flexowriter, or on punched cards. (Eventually, photo-electric scanning devices may be perfected to "read" the printed text.) The heart of the machine consists of two parts: 1) the storage unit ("dictionary") and 2) the arithmetic unit normally used in computing operations. The storage unit in certain systems may consist of a magnetic drum revolving at high speed. As the Russian word is fed in, the machine "looks up" the word in its dictionary, that is, compares it with the thousands of words stored in the memory unit, until the matching word is located. Along with each Russian word in the memory unit are stored code numbers which indicate additional information about the class and function of the word, and an English equivalent or equivalents. The coded word, together with its translation, is transferred to the arithmetic unit, where it is subjected to grammatical analysis.

The problems of grammar and syntax must be "solved" at the third step, and these may entail quite complicated procedures. Russian, for example, is a language of endings: any given noun or verb may appear in as many as twelve forms. In some way or another, the computer must deal with these endings. In some systems, it may chop off the endings, letter by letter, until the stem of the word is located in the dictionary. Thus, "vodka" might be entered in the dictionary as "vodk": the ending "a" means that the word is the subject of the verb, the ending "u" means that it is the object of the verb, etc. (for all inflected words, there is a close relation between the second and third steps mentioned above). The basic meaning of the word is determined by the stem; the function of the word in the sentence is determined by its endings, and in some instances by its position relative to other words. The task of the computer is to squeeze every ounce of useful grammatical information out of the inflected word. Certain types of information are known beforehand to be of little use in translation, and the machine will by-pass these problems, blissfully unaware of their existence.

When the grammatical information has been extracted, the English word in its modified form will be printed by devices similar to those in the input unit. If necessary, the word may be retained until surrounding words in the text have been processed, and then printed with whatever further modification the context of the sen-

tence demands. Word-for-word translations are adequate for most scientific literature, but an elegant rendering into English depends upon the degree to which each Russian word is compared to surrounding words in the sentence.

Consider the following sample product of MT, taken from a Russian textbook on mathematics. The Russian word order and punctuation has not been altered. The slant (/) gives the reader a choice of two words.

> Concept of segment known from elementary geometry: segment is part of straight line, bounded from both sides by two points. Essential characteristic of segment is its length. Recall briefly what/that is understood under this term. Take any arbitrary, but definite segment in quality of unity of length and by this segment measure given segment.

This is hardly "good" English, but it is easily understandable, especially by a mathematician. (Presumably, only persons interested in mathematics will read it anyway.) The translation can be polished to a greater or lesser degree, depending upon the demands of the consumer. The choice of two words ("what/that"), indicated by the slant, can probably be made by the machine, but only through a complicated procedure; we can also restore in translation the words "is" and "are," normally omitted in Russian. Articles ("a," "an," and "the") never occur in Russian, and their restoration appears to be impossible. Scientists will tolerate a rough translation, since they are interested only in the bare idea of the original, and since they have a clear idea beforehand what the Russian is trying to say. A smoother rendering is possible if the Russian word order is rearranged in a standard English sentence pattern. For example, the following type of construction is not uncommon in Russian: "The cited in Chapter 4 theorem of Markov. . . ." The English word order ("The theorem of Markov, cited in Chapter 4 . . .") can be restored in translation of all such constructions, if it is decided beforehand that this is necessary. In other words, a relatively smooth translation costs the machine (and its operators) a great deal of effort; there is probably a point of diminishing returns in this process, and existing digital computers are not likely to carry the procedure beyond a certain point of refinement.

There appear to be no serious vocabulary problems in translat-

ing scientific Russian. The total number of words used in the field of mathematics, for example, probably does not exceed three or four thousand. In other scientific specialties the number will not be significantly greater. Large computers can store whole dictionaries in their memory unit, perhaps enough to handle the whole field of scientific and technical literature. It may be more economical to use small-size mechanical dictionaries in translating a given specialty such as mathematics or atomic physics; when the machine tackles another field, the appropriate dictionary can be switched on.

The choice of English synonyms to translate a given Russian word is presumably beyond the powers of the machine. The mechanical dictionary would contain one cover-all English translation for each Russian entry. Russian, like English, has a certain number of words with double meanings. Thus, the word "pol" in Russian may mean either "sex" or "floor." In translating an article on biology, the latter meaning might be eliminated from the dictionary. A second possibility is to give both meanings, and let the reader take his choice. The final possibility is to instruct the machine to make its own choice, based on the nature and meaning of surrounding words in the sentence. This would involve an extremely sophisticated approximation of what goes on in the human brain, and only intensive research will determine whether or not the process can be handled routinely. At any rate, the problem of double meaning does not appear to be serious in scientific literature.

We have given in very brief form some of the problems and possible methods of MT. We should add that this research is not a "stunt," but a quite serious commercial venture for at least two industrial concerns. A substantial amount of technical literature has been composed on the subject of MT, including at least one doctoral dissertation. Within the next year or two we can expect the development of new automatic computers designed specifically for language translation; if the product is satisfactory, there will be an economic incentive for the perfection of these first models. The giant computer of today may be little more than a Model T, compared to the Cadillacs of the future.

What will be the effect of MT? In the immediate future, we can foresee the mass translation of Russian scientific literature. Suppose, for instance, that all the Russian publications on biochemistry for the past twenty years were collected in one room. The machine could produce adequate translations of the lot in a matter of

days, and then proceed to other matters. Subject indices could be prepared in advance, so that when Russian words referring to certain materials or processes were encountered in the text, a notation would be printed by the machine, giving page number and volume of the given item. Who will pay for this work? No one in the field of MT research is sure of the answer to this question. No one even knows to what use the machine will be put. It would seem that the sponsorship of the Federal Government might be required at first, since the government is now the largest consumer of Russian printed matter. But whatever the uses and the effect of mechanical translation, there seems little doubt that it will be perfected and that it will find a place in our international life.