Mechanical Translation Work at the University of Michigan

A. Koutsoudas and R. Machol, Willow Run Laboratories, University of Michigan

THE PRINCIPAL differences between the work at The University of Michigan and other work in machine translation is in the emphasis placed on the problem of multiple meaning and the approach to that problem. Our approach consists in translating small groups of words, listing in the dictionary multiple meanings under each word in the group, and finding algorithms which make it possible to choose the proper set of meanings for the group. Some 9f the dictionary meanings under each multiple-meaning word will be vacuous and some will be redundant. The algorithms are based on the pattern of vacuous translations in the dictionary for the group of words under consideration. For example, for a particular idiomatic three-word sequence, the fourth meaning under the first and third words might be vacuous, and the entire idiom will be translated under the second word. The algorithm will be such as to lead the machine to pick the fourth meaning for each word in this case. These algorithms are discussed in more detail in the article on page

Since the problem of multiple meaning cannot be solved apart from the entire problem of translation, rules are also being prepared for the syntactical and grammatical aspects of translating Russian into English, and a large corpus of Russian is being processed. At the present time 64,000 running words (128 pages) of material from the Journal of Experimental and Theoretical Physics is being coded onto punched cards, and experiments are being carried on in which technicians simulate a computer in trans - lating according to the stated rules. Theoretical frequency studies are also underway. These studies will use the results of the punched-card analysis. The theoretical aspects are based on equations comparable to those of Zipf's law. It is hoped to be able to predict answers to such questions as: How many different words will be found in a million running words? How many new words will be found in a second sample equally large? How many words must there be in a dictionary to ensure having 99% of the words in a sample randomly chosen from a certain field ?

The University of Michigan also presented to the meeting a recent idea for a Universal Font of type for technical periodical literature. It is assumed that within a generation machine translation will be a fait accompli, as will machine reading (i.e., the scanning of printed matter with the production of signals suitable for feeding a computer). All of the great mass of technical periodical literature will then be routinely translated into many languages. At that time a number of trivial problems will arise, involving differences in type faces (fonts), diacritical marks, displayed matter (e.g. equations), underlining, the use of italics or boldface to convey special meaning, etc.

When mechanical reading and translation are routine, these trivial problems will be solved by international standardization. However, this will leave the great bulk of the technical literature published in the intervening years either untranslatable or translatable only with great extra difficulty. It is therefore suggested that this standardization be performed now, so that all technical literature published after, say 1960, would be translatable by machine. As a first step it is suggested that a universal font be established. For this purpose it will be necessary to make the following studies: (1) The readability of various fonts, from the human engineering point of view (accuracy and speed) and from the publisher's point of view (appearance and reader satisfaction). (2) The machine requirements. This will involve some crystal-ball estimates as to what the finally successful reading device will be like. Of course, such machines will eventually be able to cope with certain differences, but their task will be made enormously easier if they do not have to cope with the difference between K and K or between T and T.

It may be possible to standardize also on certain other things. For example, most equations are numbered, in parentheses, at either the beginning or end of the line. It might be possible to standardize on the beginning of the line, and to use the open-parenthesis sign, (, at the left to indicate any displayed matter. This could be a cue to the machine to photograph rather than translate.

Continued on page 41