Memorandum Sept. 1963

MACHINE-AIDED TRANSLATION

Sydney M. Lamb

University of California, Berkeley

Although the Mechanolinguistics Project at Berkeley is engaged in a long-range program of basic research leading to fully automatic translation, it has found in RUSTAN (originally designed as a research tool) a practical solution to translation problems of the nearer future. At present, machine translation by itself is clearly not of sufficiently high quality to be put to practical use, nor can we expect that MT alone will be of usable quality at any time during the next five years. This means that in the meantime there must be a human factor involved.

The usual approach has been to have machine translation followed by post-editing, but this approach is futile, since post-editing of machine translation is a very arduous task and one which is more difficult than total translation by a human being. The reason that post-editing is so difficult is that (a) the post-editor must first locate the mistakes; (b) he must then determine what was said in the original Russian; (c) only then can he supply the English as it should be. The first job, that of locating the errors, is very time-consuming if done conscientiously, since the errors do not always make their presence known in the output. Sometimes one can know that a sentence has an error in the translation because it does not read right in English, but this is not true of all errors, and it is by no means easy to locate those errors which do not result in garbled English. In fact, the only way to be sure of locating all errors, as well as the only way to correctly identify what the original Russian said, is

to have the original Russian and to examine it along with the machine output. Moreover, this job can properly be done only by a skilled translator. But it is much easier for such a person simply to take the Russian and translate from it and ignore the machine output altogether.

Therefore, the only type of improvement over straight human translation that can be effective during the near future is a scheme which may be called machine-aided translation. In machine-aided translation (MAT) there is also a machine phase followed by a human phase, but the machine phase does not consist of the machine attempting to produce a translation (with the result that it has to make guesses here and there). Instead, the machine phase produces a text analysis or annotated text. Unlike a translating program, RUSTAN does not result in a loss of information with regard to the material presented to the human part of the partnership. Instead it preserves the original Russian, but segments it into its lexes and assigns as much grammatical, lexical, and other information as it can to each of the sentences, together with English equivalents, presenting the whole analysis in a neat format easily readable by a properly trained human partner. The human being, who may be called the encoder, then takes the material of the text analysis and encodes it into English.

The advantages of MAT over straight human translation are: (a) *an* encoder requires only a rudimentary knowledge of Russian, such as that attainable through about two or three years study of Russian in college; (b) such people are far more numerous than fluent Russian translators, and can be obtained at reasonable hourly rates of pay. The ordinary person hired for this type of job would have a bachelor's degree, with two or three years study of Russian and a number of scientific or technical courses appropriate to the fields in which he would be dealing. Such a person could produce translations from the RUSTAN output at a relatively high rate of speed.

A rough estimate of the cost of MAT is given in the table below. In this table it is assumed that a suitable print-reading machine will not be available at the time that a MAT system is first ready to go into production, so input is handled by flexowriter. In the table below, the cost for flexowriter and operator is given as $2.00 per thousand words, but this assumes that machine-aided translation into only one target language is being made for the source language text. It is planned, however, that before long ether target languages will be dealt with besides English. When this happens, the cost of flexowriting can be prorated among the two or more target languages. Therefore if there are two target languages, Spanish and English, then only half of the total cost of flexowriting is attributable to the translation to each of them. Similarly, if there are four target languages, English, Spanish, French, and Hindi, then the cost of flexowriting for each is only $.50 per thousand words.

ECONOMICS OF MACHINE-AIDED TRANSLATION SYSTEM

Component	$ hr.	*kilologs/hr.	$/kilolog
Flexowriter & operator	3	1.5	2.00
Converter**	25	100.0	0.25
7090	360	1440.0	0.25
Printer (1401)	60	30.0	2.00
Human "encoders"	3	1.0	3.00
Typists	2.25	1.5	1.50
			9.00 (i.e. s1ightly less than 1 cent a word)

Since the different components of a MAT system vary widely with regard to the speed of text-processing it would take a great many human encoders to keep one 7090 busy or, conversely, for a single human encoder working full-time (40 hours per week) only about two minutes per week of 7090 time is required to keep him busy. In this type of configuration the output rate would be approximately the contents of one issue of a Russian scientific journal per week, since it is estimated that a single human encoder would be able to process about one-thousand words per hour. The configuration for this rate of output is shown in the table below.

CONFIGURATION FOR OUTPUT RATE OF 40 KILOLOGS PER WEEK

Type of Component	kl./hr.	hrs./week
Flexowriter & operator	1.5	27
7090	1440.0	0.03 (i.e. 2 minutes)
Printer (1401)	30	1.3
Human Encoder	1	40
Typist	1.5	27

At 40 kilologs per avg. issue of a Russian Scientific Journal, this would be one issue per week. Russian Scientific Journals commonly have 6 issues per year (or one every 2 months), so this system would provide complete translations of 8 journals

It is planned that the RUSTAN system will be modified periodically as more and more of the basic research leading to the long-range goal of MT is completed. The Berkeley Mechanolinguistics Project has designed a machine translation system consisting of seven stages, for each one of which there is projected a program and a body of linguistic information.

The developmental work of the project has concentrated on the earlier of these seven stages and particularly on the first two. The first RUSTAN system to be used in MAT will be RUSTAN I, which will incorporate only stage one of the projected future MT system. This is the stage of segmentation and dictionary look-up. As soon as this system is debugged, stage two of the translation system will be added, and this will give RUSTAN II. Stage two will add to RUSTAN I the capability of converting from the graphemic to the lexemic stratum. Then, as soon as the project has a more complete and accurate set of tactic rules then it has at present, stage three can be added to give RUSTAN III, an even more efficient machine-aided translation system. Naturally, each time that RUSTAN is converted into a more capable system, there will be less work for the human encoder. The tables shown above give the estimates for RUSTAN I. Here, the most expensive single component of the system is the human encoder. For RUSTAN II there should be less for the human part of the partnership and for RUSTAN III, less still.

For reference, a table showing the overall projected fully automatic translation system is shown below.

OUTLINE OF THE BERKELEY TRANSLATION SYSTEM

Stage	Program	Linguistic Information	Resulting Status of Text
0			String of Graphemes
1	Segmentation and Look-up	"List" of Lexes with Segmentation Codes	String of Addresses Representing Lexes
2	Upward Conversion	Morphological Codes and Lexemic Rules	String of Addresses Representing Lexemes
3	Tactic Decoding	Tactic Codes and Tactic Rules	List Structure Representing Lexeme String and Tactic Constructions
4	Upward Conversion	Sememic Codes and Rules	List Structure Representing Sememes and Their Combinations
5	Downward Conversion	Metasememic Codes and Rules	String of Addresses Representing Metalexemes
6	Downward Conversion	Metalexemic Rules	String of Metamorphographemes
7	Downward Conversion	Metamorphographemic Rules	String of Metagraphemes

While work continues toward the goal of fully automatic translation at Berkeley, work aiming toward implementation of machine-aided translation from Russian to English Is being conducted at the System Development Corporation. In addition, the Russian-English Dictionary which has been prepared by the Berkeley project is being converted into a Russian-Spanish Dictionary by the Machine Translation Project at the National University of Mexico. Thus machine-aided translation into Spanish from Russian will also be possible, and the same Russian texts can be provided with translations in both English and Spanish. Similar dictionary conversion work by other projects to provide still other target languages is currently being planned.