[*Proceedings of the National Symposium on Machine Translation*, UCLA February 1960]

Session 11: EQUIPMENT

THE HIGH-SPEED GENERAL-PURPOSE COMPUTERS IN MACHINE TRANSLATION

B. D. Blickstein
C-E-I-R, Inc.

C-E-I-R has held an active interest in the field of machine translation since May of 1958. During this time we have programmed an MT system on the IBM 704, and are currently engaged in the design of a system for the IBM 7090. In making some observations about machine translation, we have noted in particular the considerable discussion regarding selection of computers, design of special translating machines, and requirements for the ideal dictionary device. I shall summarize some of these arguments, and attempt to arrive at some conclusions about the equipment available to the translation effort.

In order to discuss the equipment limitations imposed by machine translation, it is necessary to consider the nature of the translation process. The information flow itself is serial; that is, text flows through the computer in its natural sequence and consequently is processed in this order. The translation method determines the total number of passes that must be made through the text, and this number is an important parameter in the selection of the type of equipment to be employed. The information treated in each pass of text is different; the first pass will consist of a conversion of input text into a machine-recognizable form, and the last pass will serve to output target-language text from the computer to a printing device. Intermediate passes will serve to relate text to dictionaries, to process basic word and phrase meanings, and to perform syntactic rearrangement. As the linguistic duties of the processor are separable and may be independently performed, it is possible to break down the action of the entire system into sub-programs, each requiring one pass through the text.

Such a serial multi-pass system of organization permits continuous processing of texts of indefinite lengths, a procedure followed in payroll or accounting jobs. There are distinct differences between accounting and translating procedures, but the flow of data is similar; and, by and large, language translation presents, to the data-processing field, problems similar to many which have been previously handled with considerable success.

In the case of a payroll program, the input data consist of information identifying a man and giving the number of hours he has worked. By using the name as a key, a master file is interrogated, yielding the hourly rate of pay or perhaps an abstract pay classification. From this master-file information the program must, by various arithmetic and logical steps, determine the salary due a man, subtract the various deductions, and eventually print a check. It will be noted that once the master file is interrogated and information extracted, the employee's name itself has no bearing on the subsequent processing procedure.

Similarly, for purposes of translation, the source-language word may be completely discarded once a dictionary reference is made, if the translation can be resolved into a logical transfer of form from the target language. Miss Lukjanow has shown that such a process may even be made independent of the particular target language. The resulting translation algorithm is a logical process, involving decisions, branches, selections, and Boolean-type matches of numeric codes.

Certainly such an abstract system is independent of any particular machine, and completely indifferent to what or who performs the algorithm. However, the nature of the operations involved places certain constraints upon the choice of equipments with respect to operating economy and programming ease. Probably the major consideration is that imposed by the dictionary-search process, an absolutely indispensable task. A great deal of thought by many manufacturers and research groups has gone into this question, and a variety of possible conclusions has been examined. We shall attempt to resolve this problem by reducing the argument to two basic choices.

Inasmuch as translation appears to the computer as a multi-pass serial data-processing job, a machine chosen to do the job efficiently must possess facilities for high-speed, serial input and output. Thus far, computer technology has yielded just one device with these facilities; viz. , high-speed magnetic tape. A tried and true member of the computer-equipment family, tape is fast, reliable, relatively inexpensive, and re-usable. In addition to serving as an input and output medium, tape serves well as auxiliary storage for the computer, and is particularly well suited to serial processing. Its main drawback lies in the fact that it is not a random-access type of storage, but this disadvantage can be partially overcome by pre-sorting of files of information so that rapid serial access is available.

Recently, great strides have been made both in increasing the speed of tape units and in bringing about true simultaneity of tape-computer operation. In certain commercially available equipment it is now possible to use eight tapes simultaneously while continuing to allow the computer to perform independent logical operations. The value of such advances is reflected primarily in operating speed. For example, a process which translated on the order of 5, 000 words of text per hour on the IBM 704 can now produce 15, 000 words in the same time on the IBM 709. This improvement in tape operations has its greatest impact on the dictionary-search problem. As magnetic tape is serial by nature, dictionary search must be accomplished, in the interests of efficiency, in the following way:

(1) The tape dictionary must be sorted

(2) The text must be subjected to the same sort as the dictionary

(3) The two tapes must be matched against each other, extracting those entries corresponding to the text, and creating a third tape of these entries which is devoid of the source language

(4) This third tape must be sorted back into text-sequence order.

The property of simultaneity of tape systems now allows this sort-match-sort process to be considerably speeded, and in fact will permit large tape dictionaries to be economically feasible. Furthermore, this feature permits several technical dictionaries to be searched at no increase in operating time.

The only practical alternative to this type of dictionary-search procedure is a memory device possessing high random-access speed and capable of storing an entire dictionary at one time with equal access to any entry. In a system such as Miss Lukjanow's, each total dictionary entry should require, on the average, 20 machine-words of storage; thus a dictionary of 50, 000 entries would require 1 million words of storage. It is estimated that at least 100, 000 words per hour can be translated through a tape-equipped IBM 7090; a comparable random-access memory would then need to have an average access time of 36 milliseconds for an entire entry, exclusive of processing time. Thus the transmission rate from storage would have to be on the order of 1. 8 milliseconds per word. This is considerably faster access than has been available in past units of the magnetic-disc type, and consequently we are led to choose magnetic tape as the best currently available dictionary medium. It is expected that the near future may bring about disc files that meet and even better this access time. The STRETCH disc file will have an access time on the order of 4 microseconds per word, but this presupposes that the proper track on the disc has been located, an operation which may require up to 87 milliseconds. Actually, a great deal of this lost location time can be recovered by proper buffering and interweaving of processing with lookup.

One of the questions we are attempting to answer is whether current machines have been designed to meet the needs of linguistic material as opposed to numerical material. I should like to cite the chain of development of the IBM 700/7000 series. The first of this group was the 701, which contained basic arithmetic orders and very little else. It was soon followed by the 704, which added indexing features and certain logical operations. The 709 provided 6-channel input and output, indirect addressing, and a family of well over 20 logical operations; the 7090 increased the 709's internal speed by a factor of 5, tape speed by 4. 5, and increased the channel capacity to 8. The net effect from IBM 701 to 7090 has been the implementation of logical and information-handling abilities, rather than the improvement of arithmetic facilities; the result is computers which are just as able to handle data-processing applications as they are to do numerical manipulation. There is plentiful evidence among computer users that this is so, and that the general-purpose machine has become general purpose in scope as well as in name. It has also become evident that although the information in many applications is alphabetic or decimal, the binary machine has proved itself more efficient than its decimal counterparts as a result of its higher internal speeds. Actually, the term "computer" is fast being replaced by the more appropriate name, "electronic data-processing machine".

The operations needed to perform translation are basically logical, consisting of logical "or" and logical "and", equality tests, and decision branching. Recently certain binary machines have been announced which will be capable of utilizing magnetic disc-file memories. The IBM 7090 is currently available with a limited disc file, and the forthcoming STRETCH, as we have previously mentioned, will be equipped with a disc file of extremely high capacity and access speed.

We have mentioned several general-purpose computers; and for the sake of completeness, and in order not to appear to be biased toward IBM, we should like to give a brief summary of a number of these general-purpose machines.

Despite the preference for binary-type machines, there are some relatively new and large-scale decimal models on the market today which merit consideration. The Univac LARC, the Datatron 220, and the IBM 7070 are all machines of extremely high speed for their class, and all of them have been equipped with a set of logical-type instructions. Disc files are available for the 7070, and all three are equipped with tape units of high speed. Among the binary machines are two basic groups; those with vacuum-tube circuitry, and those with transistor logic. In the former group are the IBM 704 and 709; in the latter, with operating speeds of about 5 times the vacuum-tube speeds, are the Honeywell 800, CDC 1604, IBM 7090, and Transac S-2000. It is felt that the transistorized group offers the best buy available for translation, as the speeds are achieved at a cost increase of only about 50% over the vacuum-tube types, yielding an over-all reduction in cost of about 70% on each word of translation.

I have mentioned several times a machine called STRETCH. This computer, designed by IBM for the Los Alamos Laboratory, is a "stretch" into the future both electronically and logically. Operating speeds internally are about 100 times faster than those on the 704, and the advances made in logical design and asynchronous instruction structure will permit a 20% reduction in the number of instruction steps needed. The machine will be capable of handling a maximum of 32 disc files each of capacity of over 4 million machine words, for a total disc-file capacity of over 130 million machine words. High-speed magnetic tapes with a transmission rate of over 10, 000 words per second will operate in parallel with the machine and with each other; the main magnetic-core memory of up to 262, 000 words of 64-bits each will be available. Because the elimination of tape-dictionary search yields roughly a 50% increase in the translation rate, and because of the attendant increase in internal speed, it appears that a system such as Miss Lukjanow's could be implemented with a translation rate in the neighborhood of 3 million text words per hour.

A great deal of discussion, has concerned special-purpose translating machines, and in the past few years several designs have been advanced, and some actually built. The dangers of this are evident if one notes the rapid obsolescence which is characteristic of both computer technology and advances in language methods. By contrast, the user of the general-purpose machine may rent his equipment from the manufacturer, thus saving a costly initial investment and may, furthermore, always avail himself of the newest and most efficient equipment.