[*Times Literary Supplement*: Freeing the mind, April 1962]

4—THE KINDS OF MACHINE NOW IN USE

By Andrew D. Booth

BEFORE ATTEMPTING TO DISCUSS the types of machine which are available to act as ancillaries to human effort in the fields of linguistics and information processing it is necessary and worthwhile to speak briefly of the way in which data are presented to such machines—that is, to discuss coding.

The first hurdle which the linguist encounters when faced with the problem of mechanizing his art is to see how a machine which is designed for handling numbers can possibly handle linguistic symbols. In fact, this is simple and depends on the schoolboy trick in which the alphabetic characters are represented by decimal numbers—for example, A = l, B=2, . . . Z=26. The precise coding for any particular machine varies, but there exist two classes of code, the first devised originally for telegraph transmission, the second for punched-card business machines.

In the telegraph code use is made of the so-called "binary" notation in which the only digits which can occur are 0 and 1. In this scale the decimal digits have the binary equivalents shown in the following table:—

0=00 00000
1=00 00001
2=00 00010
3=00 00011
4=00 00100
... ...
9=00 01001

The way in which the binary equivalents of larger numbers are generated is easily seen. In the case of the alphabet the first convention is the one mentioned above—that is, A=l, B=2. But representing these in the binary scale we find

A =01 00001
B=01 00010
C=01 00011
. . . . .
Z=01 11010

The reason for introducing the 1 in the second position is that it enables the machine to distinguish between the binary *numbers* in the first table, and binary code numbers representing alphabetic symbols shown in the second.

Up to quite recently, when ordinary telegraph apparatus was all that was available, only five groups of punching were available for each letter, and this made necessary the representation of alphabetic symbols by a binary code preceded by a code impression to tell the machine that it was dealing not with numbers but with coded letters.

Binary coding is the most efficient which is possible in existing machines. The punched-card machine industry has developed its own code, which is quite distinct from the telegraph code shown above. In this the punched-card has a single punching in each column where it is intended to represent a decimal digit, but a pair of punchings to tell the machine when an alphabetical symbol is intended. It will be clear that the binary code, because it makes use of all of the possible positions, is a more efficient one than the punched-card code, in which only two out of twelve possible positions are in fact punched.

The next consideration is the basic types of machine which are available for the work envisaged. The first class of machines is composed of punched-card calculators designed originally for the operations of business accounting. From the point of view of the linguist and information processor, the machines can be divided into four types:

1. machines which sort
2. machines which collate
3. machines which punch
4. machines which print.

As far as sorting is concerned the machines take information, linguistic or otherwise, punched on to cards, and by an ingenious electrical or electronic system arrange that all cards which have the same punching in a given column are sent to the same "bucket" in the output of the device. Thus, for example, all cards punched with 1 in the first column are distributed into bucket 1, all cards punched with 2 into bucket 2, and so on. Naturally, if 1 is the code number of the letter A, this means that all cards bearing words beginning with the letter A appear in bucket 1, all cards beginning with letter B, appear in bucket 2 and so on.

Collating machines are less common than sorters; they have the desirable property of being able to take two separate sets of cards, to pass these through two feed mechanisms and to distribute them between a pair of output buckets according to the satisfaction or otherwise of certain criteria read by the machine from the cards under the reading stations at the given instant.

Having dealt with the cards in this way, that is either by sorting or collating, it is necessary, unless purely statistical information is required, to produce a typed or punched version of the result. This can be done either by punching results on summary cards, or by the direct printing of the output on a printing tabulator. The essential feature to remember in connexion with these machines and also with the more elaborate electronic versions of them which, besides sorting and comparing, can also perform the operations of elementary arithmetic, is that the machines generally have no appreciable internal storage. That is, it is not possible to take all of the data which are contained on, say, 10,000 cards, and hold them in reserve within the machine while performing some other operations.

On the other hand this class of machine is relatively inexpensive and is available in many ordinary offices. For this reason they have been attractive to non-commercial users who profit from the spare time which is occasionally available either free or at greatly reduced costs. As far as speeds are concerned, as a general rule sorting and collating take place at speeds which vary from 200 to 600 cards per minute. Punching and printing on the other hand are slightly slower, from say 200 down to 100 cards per minute. The important thing to realize is that each punch or print operation takes place simultaneously, over between 40 and 120 columns on each card, which means that a line of up to 120 letters can be printed at one operation.

The second basic type of machine is the electronic computer. In effect this consists of circuits, derived originally from the radar of the Second World War, which enable the operations of addition, subtraction, multiplication and division to be performed at speeds which range from 1,000 to 1,000,000 per second. Associated with these arithmetical units are first and foremost the electronic equivalent of pencil and paper, usually referred to as the store.

This often consists of a hierarchy of distinct units ranging from a relatively small but very high-speed inner store to an almost infinitely large, but very slow, outer one.

The third unit of the machine is the control, which accepts instructions in coded form from the store and causes their execution by other units of the machine and is also capable of making decisions in accord with the results of calculations already performed, and of ordering the operations of the machine in the light of these decisions. Finally, the machine must have some form of input from the outside world and also an output to communicate its results to men.

From the linguist's point of view the most important features of the machine are the input and output since these are the units by which he makes contact with what otherwise amounts to a large black box. Input, as previously mentioned, usually takes the form of punched tape or punched cards. A form of input which is of potential importance, both in data processing and in linguistics, is the Monotype rolls created during the course of printing. At least one such Monotype reader is now in existence and there is no reason in principle why any computing machine should not have access to material set up in this medium. One disadvantage of the method is, however, usually forgotten; it is that the Monotype roll is the initial record of a rather complex operation. Generally speaking it is not corrected in the light of proof readers' comments, neither does it contain an adequate representation of the complex settings required in mathematical typology. Nevertheless, for non-mathematical material the Monotype roll may in future form a valuable computer input.

Of more spectacular interest among input devices is the direct reading of printed or typewritten characters. Here several practical devices are already in existence both in this country and in the United States. It is only fair to say, however, that, whilst these devices will read accurately from such things as the till-roll produced by a cash register, so far none has means of absorbing information presented from a page of a printed book. In the latter case the major problem, that of handling the paper without dissecting the book, has not yet been solved. As far as speed is concerned, the direct recognition of printed characters can take place at rates which vary with the particular machine from 50 to 1,000 characters per second.

Finally, in the field of input mention must be made of the possibility of recognizing the spoken word. Up to the present the mechanical recognition of speech has been demonstrated only for restricted ranges of sound such as the decimal digits 0 to 9. However, in principle, the problem of speech recognition is well on the way to solution, and it may be that some day a spoken word recognizer will form the input to a machine for simultaneous multilingual translation.

For output the same sort of devices are available; punched tape and cards are standard on many machines, and may be the only form of output provided; the cards and tapes being later interpreted by a distinct and often remotely situated reading device which prints out the results. This so-called "off line" operation is nowadays favoured, since the operation of punching is usually faster than that of printing. Jt would be possible in principle to set type directly by Monotype, but so far as the author is aware this has yet to be done.

Of more recent output processes mention must be made of xerography—that is, the production of markings on paper from images presented to it by a cathode-ray oscilloscope. Here the paper requires some form of development, usually by heating, and in consequence has the disadvantage that it cannot rapidly be started and stopped, since time must be allowed for the heating apparatus to cool off lest a fire result.

Finally, practical schemes for producing spoken words from computers have been demonstrated in several laboratories. These are off-shoots of investigations aimed at economizing the expensive time involved in transatlantic telephone conversations.

The central problem of computing machines at the present time is that of storage. It has been mentioned that machines have a store consisting of a number of discrete sections. Thus a typical machine may have an almost infinite store on punched cards or punched paper tape, and a secondary store on magnetic tape of just the sort which is familiar in tape recorders except that, instead of storing spoken words, it stores the coded impressions mentioned above. Magnetic tapes may be accompanied or replaced by some form of "random access" storage, such as the "juke box", "carousel", or "file drum". To make clear the differences between tape and these random access devices it may be remarked that, to locate a piece of information on a magnetic tape, the tape must be started and wound past the reading station until the information is located. This will take, in general, about half the time required to wind the tape from end to end and may amount to several minutes. Random access devices, however, may be likened to a stack of gramophone records. Here the first operation is to select that particular record upon which the information is to be found, and the second to locate the band or track which more specifically refers to the desired data. It is fairly clear that since the gramophone arrangement allows quite fast access to any particular record without sorting through the whole, it is intrinsically faster than the tape. Typical location times for random access memories range from one-third to one-tenth of a second.

In addition to tapes or random access devices most modern machines have a large store on a magnetic drum. This is a cylinder coated with magnetic material and read by a number of parallel reading stations, something like a very wide magnetic recording tape. The advantage of the magnetic drum is that, speaking in terms of ordinary language, it is possible to store something of the order of 100,000 words and to reach any one of these in a time of about 1/100 second. Finally machines have a relatively small, but extremely high speed, store which their arithmetical unit can use. This device often takes the form of small rings of magnetizable ceramic material or of very thin films of magnetizable metal, and is capable of accepting or giving out information in times of from one millionth to one ten millionth of a second.

A previous author has mentioned that the disadvantage of magnetic tape is that the location of a piece of information involves, in general, scanning half of the store, It is only fair to remark that this is a very limited view of the problem since what is usually done with tape stores is first to sort out questions from a number of sources which are to be posed to the tape and arrange these in alphabetical order. This accumulation of questions is then compared with the tape. Thus, in one passage through the tape, very many questions can be asked, and it is possible to show that, given a sufficiently large number of questions, such a tape device is more efficient than a random access device of the speeds currently available.

Nevertheless, tapes have one disadvantage which is implicit in the remark made by Foskett, quoting Bernal, "that the best way to acquire information is over a glass of beer". The logical *raison d'être* for this statement is simply that it is trivially possible to ask a fellow human being whether he knows the answer to a question and to obtain from the answer yes or no. This cannot be done with any of the types of computer store mentioned above, since the answer to a question involves looking at all of the information one item at a time. The late Dudley Buck, however, devised a new form of store making use of the phenomena of super-conductivity which take place at temperatures near to the absolute zero. This so-called "cryogenic store" has the interesting property that it can answer the question "is the information available?" without at this stage actually locating it. The importance of this is that reference can be made to large areas of storage to find out whether it is worth scanning them at all, so that using a cryogenic store a detailed scan would only be necessary in that portion known from its own statement to contain the required information. This form of storage is very likely to produce a revolution in computer technology as well as in ideas for computer programming during the next decade.

Finally, some remarks on the applicability of computers in the fields of linguistics, documentation and information processing. As a literary aid the computer is already well established. Glossaries and concordances have been constructed and the computer has been used to establish chronological dating in the Platonic dialogues, while at the present time stylistic studies by computer are directed to investigating the authenticity of the Pauline Corpus. Demonstrations of machine translation are almost daily occurrences although, in the view of the present author, which may be to some extent authoritative since he was the inventor of this science, machine translation as such is not of great practical importance.

Machine translation research may be of value, however, in that, because one of its basic techniques is the establishment of a meta-language or intermediate language of ideas, it is in principle capable of constructing automatically an abstract of the important features of a given text. Furthermore, machine translation-like procedures may have important applications in the whole field of information retrieval. In particular, although much play is made of the statement that such things as Chemical Abstracts have expanded by a factor of three over the last decade, nothing is said of the expansion of the original or worthwhile content of the papers so abstracted. In fact in a survey of the computer field in 1960 the author discovered that about 10,000 pages of published literature could be represented adequately so far as all new ideas are concerned by forty pages of descriptive material.

This reduction of large amounts of received material to a small corpus of really relevant facts may soon be achieved by machine translation techniques and one view of the information retrieval centre of the future is that a large machine will contain in coded form all of the original material available on a given subject without any of the actual authors' words. On presentation of a question all of the relevant information will be extracted and, using some of the techniques of random sentence construction, an account of the field in respectable English will be produced. Naturally, such a machine will not be popular among authors since by and large no reference will be made to the originator of any given idea. But, on the other hand, perhaps this will be all to the good since it will avoid the hard feelings which are so often produced at the present time by inadequate referencing.

The Times Literary Supplement—April 13, 1962