CALCULATING MACHINES AND MECHANICAL TRANSLATION

ANDREW D. BOOTH
D.Sc., Ph.D.. F.Inst.P.

As far as can be traced, the first suggestion that translation —from foreign language (for which the letters F.L. will afterwards be used as a convenient abbreviation) into a target language (T.L.)—could be achieved by means of a digital calculating machine was made by the present author in 1947. The scheme, as it was then envisaged, merely consisted of using the storage unit of a modern computer to hold the contents of what amounted to a dictionary. This idea was developed in the following years in collaboration with R. H. Richens. and has now emerged as an entirely satisfactory method, at least for the translation of scientific texts.

A number of workers in the U.S.A. have taken up the development of the method and, with their superior resources and the backing of certain U.S. government agencies, they have made considerable progress towards the setting up of installations for the practical application of the method.

The general interest in the matter was shown by the holding—under the auspices of the Rockefeller Foundation—of an international conference on mechanical translation. This took place at the Massachusetts Institute of Technology in June 1952, and a symposium volume, based upon the proceedings, is due to appear in the near future.

It is thus appropriate to set down, briefly, the nature of the machines and the processes involved, and to indicate the possible range of the method.

MODERN DIGITAL CALCULATORS

Since the operation of a scheme for mechanical translation depends upon techniques which are a part of the modern development of electronic digital computers, it will assist the reader to understand the methods used if a brief account of the relevant features of such machines is given here.

In essence, a computer of the type under discussion consists of four parts:

(i) A unit for performing the operations of arithmetic.
(ii) A control unit for directing the operations of the rest of the machine.
(iii) A store for numbers and instructions (the 'memory').
(iv) Means for inserting data into the machine and for indicating the results of its operations (input/output).

It is not necessary to go into further detail as to the precise functions of these units, except to notice that the store of a modern machine is usually capable of holding many thousands of numbers.

The reader may be under the impression that these machines are necessarily of enormous size and complexity. That this is not the case is shown by the photograph of the A.P.E.R.C.—the All Purpose Electronic Rayon Computer which is working at Birkbeck College, London. This is probably the most compact machine at present in operation, and the extensive planning which went into its design has been amply repaid by the reliable functioning of the installation over long periods; it has, for instance, given 93% faultless operation over a period of 390 hours.

Before leaving the subject of computers, it is necessary to mention that most of these machines work in the so-called binary scale which contains only the digits 0 and 1. It follows that numbers in normal decimal scale have to be represented as binary equivalents, in a manner which will be seen from the following table:

TABLE 1

Decimal	*Binary*
0	0
1	1
2	10
3	11
4	100
5	101
6	110
7	111
8	1000

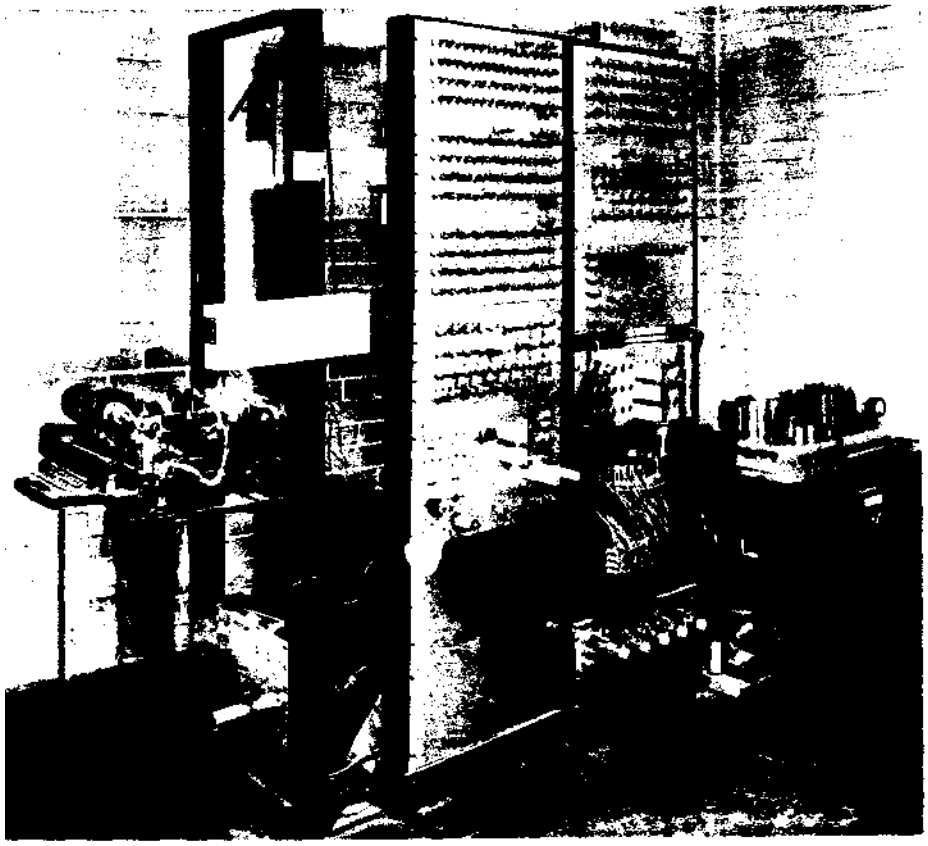
etc.

LANGUAGE AND NUMBER

To be able to perform a translation on a computing machine—or indeed on *any* machine—one must be able to represent words by means of numbers. A familiar example of this occurs in the ordinary printing telegraph where the depression of a key on a teleprinter has the effect of transforming the given letter into a stream of electrical impulses, spaced out in time. The international teletype code, for the alphabet, is shown in Table 2, where the symbol ○ indicates an electrical impulse of one type whilst ● is either no impulse at all, or else the negative of ○. It is at once evident that this code is essentially a binary one and that possible 'numerical values' of the various letters are those given in the third column of the table.

TABLE 2. TELETYPE CODE

A	●	●	○	○	○	24	N	○	○	●	●	○	6
B	●	○	○	●	●	19	O	○	○	○	●	●	3
C	○	●	●	●	○	14	P	○	●	●	○	●	13
D	●	○	○	●	○	18	Q	●	●	●	○	●	29
E	●	○	○	○	○	16	R	○	●	○	●	○	10
F	●	○	●	●	○	22	S	●	○	●	○	○	20
G	○	●	○	●	●	11	T	○	○	○	○	●	1
H	○	○	●	○	●	5	U	●	●	●	○	○	28
I	○	●	●	○	○	12	V	○	●	●	●	●	15
J	●	●	○	●	○	26	W	●	●	○	○	●	25
K	●	●	●	●	○	30	X	●	○	●	●	●	23
L	○	●	○	○	●	9	Y	●	○	●	○	●	21
M	○	○	●	●	●	7	Z	●	○	○	○	●	17



THE A.P.E.R.C. MACHINE AT LONDON'S BLRKBECK COLLEGE

When a number of keys is depressed in succession the result is the emission of a succession of impulses which, in aggregate, give a number which is characteristic of the word typed, thus:

and =11000, 00110, 10010=24,786

The crudest way in which a calculating machine with storage facilities could perform the operation of translation can now be indicated. Suppose that each storage position* is given a number, then in the storage position bearing the number which represents the word to be translated a further number is placed. This new number gives, in coded form, the translation of the first word. Thus, suppose we are translating English into French; for the word *and,* we find in storage position 24,786, the number 513 or, in binary 10000, 00001. The reader will see, on examination of Table 2, that this represents the letters *et* and by presenting the coded number to a teleprinter the French translation could be printed out.

This simple example serves to demonstrate two things: firstly that translation can be effected by means of an essentially arithmetical machine, and secondly that this particular method is quite impracticable. To emphasise the latter point, it may be remarked that the code numbers of all words of not more than 10 letters would lie in a numerical range whose greatest member is of order 2^{50}—or about 10^{15}. No computer storage device is conceivable which has anything approaching this capacity; in any case, the total number of words in any language is probably less than 10^7 so that only a fraction (10^{-8}) of such a store would be occupied.

Fortunately, there are other and more reasonable methods of approaching the problem: one of the simplest is to store the numerically coded form of the F.L. word *and* the code number of the translation in the same storage position. The computing part of the machine now separates off the F.L. digits of each stored word group in turn and compares these, by subtraction, with the digits of the word to be translated. In general, only one of these results will be *zero,* and this corresponds to the correct translation position; modern computers have means of detecting, either directly, or by means of an inserted order sequence, the nullity, or otherwise, of a number so that nothing new is involved. All that remains, is for the machine, having recognised the identity of the F.L. part of a particular entry with the F.L. word to be translated, to output the remainder of the composite entry, which is the required translation.

Even this simplified scheme would require a considerable storage capacity to translate a reasonable proportion of the words occurring in, let us say, a scientific paper, and it was left to Richens to suggest a method which reduced the problem to one well within the capacity of existing machines. This technique will be discussed in the next section, but before leaving the subject of codes for alphabetic characters it is worth mentioning that, for simplicity of

* A modern digital computer has means of storing some thousands of numbers which result from its calculations. These may be likened to the lines on a sheet of paper, and each line is numbered 1, 2, 3 … etc. The computer can execute commands of the type "Read the number on line 625" or, "write the answer on line 720".

calculation, the teletype code given in Table 2 has never actually been used in a computer. Instead, the letters art numbered in ascending order of magnitude, thus:

TABLE 3

A=00001=l
B=00010=2
C=00011=3
.....
Z=11010=26

This is easily arranged on the teleprinter and simplifies considerably the comparison process used in the machine.

MICRO-SEMANTICS

The practical application of machine translation depends upon two things: firstly that the different words which occur in scientific papers on a given subject are limited in number, and secondly that a simple means exists for dealing with the variants introduced by stem-ending combinations.

Limitation of vocabulary is familiar to anyone who reads specialist papers in a foreign language; thus the set of words required to translate a mathematical paper is quite different from that needed in, say, genetics or brain surgery. Richens and the present author working in this country, and Oswald and Bull (1953) in the U.S.A., have examined this problem, and the conclusion emerges that, to translate a large proportion of the words in a scientific text, some 1000 words of specific scientific application, together with a similar number of words of general literary usage, will suffice. This limited vocabulary has been termed a 'micro-glossary'.

A normal dictionary does not contain an entry for each possible variant of a particular word; thus the word *calculate* will appear, but not *calculates, calculating, calculated* and so on. Richens has pointed out that by 'storing' the stems and endings in a dictionary separately a much more useful output could be obtained from a small number of entries than would otherwise be the case. For example, the verb stems:

calculat-
differentiat-
lov-.
mat-
not-
releas-
teas-
undulat-
violat-
wak-

can all take the common endings *-e, -es, -ing -ed* when used as verbs, the endings *-er* and *-or* when used as nouns and so on. Thus by storing stems and endings separately in the cases just quoted, 10 stems and 6 endings would enable 50 to 60 different complete words to be translated. The way in which this is put to use in a calculating machine, or other mechanical translator is as follows: the complete F.L. word is first coded into binary numerical form. The number thus resulting is then compared, by subtraction, with the F.L. part of each dictionary entry; starting with that part of the dictionary which contains complete words. It may so happen that the exact entry is found, in which case the translation is produced at once; if, on the other hand, no exact equivalent is present, one letter (or more accurately the group of digits representing one letter) is removed from the end of the F.L. text word and the process is repeated. In this way a point will be reached at which the longest portion of the text word, corresponding to a stem entry in the 'dictionary', is found. When this occurs the stem translation is printed out, and the letters (in coded form) which have been removed are examined by comparison with the ending dictionary contents. The result of this final comparison then appears as 'grammatical notes' in which, for example, the person and tense of verbs are indicated.

This description gives, of course, a simplified picture of the complete process that is needed to ensure translation of the multiple words in a language such as German, and to take care of multiple word units such as *a-t-il* and *ne … pas* in French, but it should help the reader to grasp the general manner in which an essentially arithmetic machine can deal with a problem which, at first sight, appears to have little to do with calculation.

It may be of interest to see three specimens of the sort of output which is produced by such mechanical translation. The first is a translation of an Italian piece from the literature of plant genetics.

Original passage in Italian

E' stato prov/ato che i cereal/i d'invern/o cresc/iuti in serra mostr/ano poc/a resistenza al freddo, mentre gli stessi cresc/iuti in campo apert/o, sono molto/o piu resistent/i.

Machine output

is [been / status] prove (p) [that / which] (v) cereal (m) of winter (z) grow (pm) in [mountain / crowd / greenhouse] show (m) little (v) resistance to cold while (v) same (m) is (ps) grown (pm) in field open (v) are much (v) more resistant (m).

English translation

It has been proved that winter cereals grown under glass show little resistance to cold, while those grown in the open are much more resistant.

It should be noticed that the grammatical notes supplied by the machine are:

(m) = multiple or plural
(p) = past
(s) = subjective
(v) = vacuous, i.e. having no English significance
(z) = unspecific
/ = stem-ending separation point.

The second example shows the result of supplying a message in Russian to the I.B.M. '701' data processing machine. Some of the stages involved are shown opposite in Table 4.

TABLE 4

How the passage is analysed

RUSSIAN WORD	ENGLISH EQUIVALENTS I	II	1st CODE	2nd CODE	3rd CODE	RULE NO.
vyelyichyina	magnitude	---------	***	***	**	6
ugl-	coal	angle	121	***	25	2
-a	of	---------	131	222	25	3
opryedyelyayetsya	is determined	—•-----	***	***	**	6
otnoshyenyi-	relation	the relation	151	***	**	5
-yem	by	---------	131	***	**	3
dylin-	length	---------	***	***	**	6
-i	of	---------	131	***	25	3
dug-	arc	---------	***	***	**	6
-i	of	---------	131	***	25	3
k	to	for	121	***	23	2
radyius-	radius	---------	***	221	**	6
-u '	to	---------	131	***	**	3

What the Rules mean:

RULE 1:

REARRANGEMENT

If the first code is 110, is the third code associated with the preceding complete word equal to 21? If so, reverse the order of appearance of the words in the output (i.e. a word carrying 21 should follow one carrying 110)—otherwise, retain the order.

In both cases, English equivalent I associated with 110 is adopted.

RULE 2:

CHOICE, FOLLOWING TEXT

If the first code is 121, is the second code of the following complete, subdivided or partial (root or ending) word equal to 221 or 222? If it is 221, adopt the English equivalent I of the word carrying 121. If it is 222, adopt English equivalent II.

In both cases, retain the order of appearance of the output words.

RULE 3:

CHOICE, REARRANGEMENT

If the first code is 131, is the third code of the preceding word or either portion (root or ending) of the preceding subdivided word equal to 23? If so, adopt English equivalent II of the word carrying 131 and retain the order of words in the output. If not, adopt English equivalent 1 and reverse the order of appearance of words in the output.

RULE 4:

CHOICE, PREVIOUS TEXT

If the first code is 141, is the second code of the preceding complete word or either portion (root or ending) of a preceding subdivided word equal to 241 or 242? If it is 241, adopt the English equivalent I of the word carrying 141. If it is 242, adopt English equivalent II.

In both cases, retain the order of appearance of words in output.

RULE 5:

CHOICE, OMISSION

If the first code is 151, is the third code of the following complete word, or either portion (root or ending) of the following subdivided word equal to 25? If so, adopt English equivalent II of the word carrying 151. If not, English equivalent I.

In both cases, retain the order of appearance of words in the output.

RULE 6:

SUBDIVISION

If the first code associated with a Russian dictionary word is … then adopt English equivalent I of alternative English language equivalents, retaining the order of appearance of the output with respect to the previous word.

Original passage in Russian

Vyelyichyina ugla opryedyelyayetsa otnoshyenyiyem dlyini dugi k radyiusu.

The message is punched upon a card and processed in various pieces of equipment shown in Fig. 2. Eventually the output—"Magnitude of angle is determined by relation of length of arc to radius"—appears on the typewriter. This particular system was evolved by Dr. Leon Dostert, who appears as the central figure in the fourth photograph on p. 284.

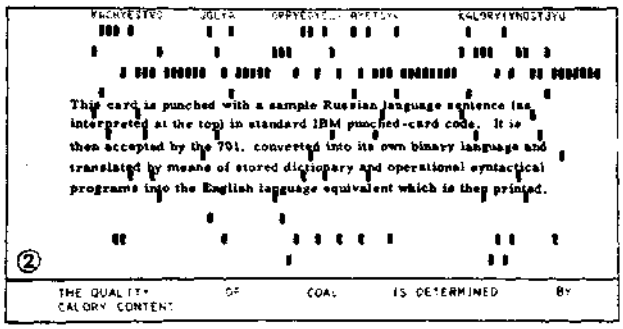
As a final example, the following short passage (from a German text on neurosurgery) is the result of the application of a statistical-frequency analysis to the word forms contained in a selection of the literature. It is quoted from one of the most recent reports of Oswald and Lawson (1953) who have been studying the application of the Institute for Numerical Analysis Computer to the problem.

Original Text in German

Die Verletzungen im Bereich der vorderen Schördelbasis stellen den Chirurgen immer wieder vor die Frage, in welcher Form derartige Verletzungen wegen der möglichen Beteiligung der Nasennebenhöhlen (NNH) am besten zu versorgen sind. Hierbei spielt die Drainage des NNH—Gebietes eine wichtige Rolle.

Machine output

The injuries in-the region of-the anterior cranium base put the surgeon always again before the question, in



The Type '701' electronic data-processing machine of International Business Machines Corporation translating from Russian into English.
1. The typist prepares a punched card. 2. Specimen punched card, and below it a strip with the translation. 3. General view of the machine. In the centre of the units shown is the Electronic Analytical Control Unit, and at its right is a Card Reader; behind the control unit is the Power Distribution Unit. On the left are the Magnetic Drum Storage Unit and the Electrostatic Storage Unit. In the right-hand group are two Magnetic Tape Readers and Recorders, the Alphabetical and Numerical Printer and the Card Punch. 4. Dr. Leon Dostert *(centre)*, chairman of Georgetown University's Institute of Languages and Linguistics, looking at continuous sheet of English-worded sentences coming from the printing mechanism. 5. Sheet of English-worded sentences produced by the printing unit of the '701'.

which form such injuries on-account of the possible participation of-the nose sinuses (NNH) at the best to treat are. Here plays the drainage of the NNH. —area an important role.

English

Injuries of the base of the anterior cranium, always place before the surgeon the question of the best treatment, in view of the possible participation of the nasal sinuses. The drainage of the nasal sinus area here plays an important part.

FUTURE PROSPECTS

Several important obstacles still remain to be overcome in the field of mechanical translation. First is the need for an adequately fast input and output for the machine; this is obvious when one considers that the translation of a 1000-word text by machine takes between 2 and 7 hours with currently available facilities, which is considerably longer than the time required by a skilled human translator. Two attractive possibilities exist for improving this situation; in the first a scanning device 'reads' the actual text or typescript to be translated; in the second the machine operates from a spoken input. Both of these suggestions are at present under investigation, the first by D. Shepherd and his group in the United States, and the second at Bell Laboratories and also at Birkbeck College. The problem of output is not so critical as a number of high-speed printers are already available. Even with the present limitations the process is still useful, since a range of foreign languages can be translated with the same equipment, whereas few laboratories would have a 'range' of multilingual humans.

A second need is for the construction of a special machine for mechanical translation since although a general-purpose digital computer can perform the operations, it is too complex in the arithmetic sense and too small in the size and quantity of the numbers of words which it can store. As an indication of the sort of thing meant, a typical existing computer 4 can store 1024 'words' each of 32 binary digits and can perform the operations of +, −, ×, and shift, whereas what is required is a machine with storage for 4000-8000 'words' of 250 binary digits and an arithmetic unit which need only subtract and shift.

On the linguistic side much remains to be done; adequate microglossaries do not exist—or at any rate are not generally available—and, it will be some time before really comprehensive stem-ending compilations are ready. It may be that the suggestion of E. Reiffler (1952) of pre-editors and post-editors will prove necessary. The pre-editor, a native in the F.L. who need not have any knowledge of the T.L., removes syntactic and morphological ambiguities from the original text, whereas the post-editor renders the machine output into respectable English (or, of course, any other T.L.). Yet again, it is possible that the suggestion of Dodd (1952) for universal scientific publication in some standardised literary form may find favour.

This field is an expanding one, in which new ideas both linguistic and engineering are constantly arising. If we cannot, at present, translate by machine the German of Goethe into the English of Wordsworth, it is by no means certain that this will be true in five or ten years' time.

REFERENCES

Booth, A. D., *Report to the Rockefeller Foundation on the proposed London electronic computer,* Princeton, 1947.

Richens, R. H. and Booth, A. D., "Some methods of mechanised translation", Proceedings of Conference on Mechanical Translation. (These proceedings are due to be published this year by John Wiley, the American publishers.)

Oswald, V. A. and Lawson, R. H., *An ideoglossary for mechanical translation.* University of California, 1953.

Booth, A. D. and Booth, K. H. V., *Automatic Digital Calculators,* Butterworth, London, 1953.

Reiffler, E., *Studies in mechanical translation,* Nos. 1-7, University of Washington, 1952.

Dodd, S. C, "Model English for Mechanical Translation", Washington Public Opinion Laboratory (Mimeographed), 1952.