[From: **Babel**, vol.2, no.3, October 1956]

ANDREW D. BOOTH

Present Objectives of MT Research in the United Kingdom

Dr. Booth is the Director of Birkbeck College Research Laboratory at the University of London and directs there an Electronic Computer Project of which he has kindly consented to provide a first-hand account of its present stage and objectives.

It is the purpose of this short note to introduce readers to the basic ideas of machine translation. This is necessary if they are to appreciate the paper *by L. Brandwood.* (See following article)

The idea that translation might be possible by means of an automatic computing machine arose during conversations between the present author and Warren Weaver, of the Rockefeller Foundation, which took place in New York in 1947. From that time until l950 very little practical work was done on machine translation, although the present author and *Richens* of the University of Cambridge worked out a tentative scheme whereby low grade translation could be effected using standard punched-card machines. A trial was made and the method was shown to be satisfactory, but it was not thought that much practical use would attend its wider development.

In the early nineteen-fifties workers in the United States became interested in the subject and this led, in 1953, to a conference which was held in the Massachusetts Institute of Technology. Following this conference a demonstration of machine translation was carried out by the I.B.M. Corporation in conjunction with Georgetown University and *L. Dostert* in which selected sentences from Russian were rendered into English. Much excellent work has been done in the United States, and although it is invidious to single out any particular investigation, the paper of *Oswald* and *Fletcher* (1), who have produced an analysis of the German language in a form suitable for an automatic calculating machine, is noteworthy.

In England work is being sponsored by the Nuffield Foundation and since 1955 my own laboratory has been engaged in producing practical schemes for rendering one language into another. To avoid duplication of effort, and to avoid competition with American projects, a modest programme was initiated to render scientific French into acceptable English. The paper by *Brandwood* will show the extent to which this objective has been realised. We consider that we are now able to translate a French scientific article having a limited vocabulary on the APEXC calculator (2). This restriction can be overcome in two ways, either by constructing micro-glossaries for the particular type of French scientific literature which is under examination, or by extending the dictionary which is stored in the machine. Both of these methods are receiving attention, the first by means of linguistic assistants who read the literature of particular scientific subjects and make dictionaries of the words which occur therein, and the second by our engineering programme which aims to extend the storage capacity of APEXC to 64,000 words.

We do not think that APEXC, or any existing calculating machine will ever be used for economical machine translation, and we are chiefly concerned with reducing the principles on which such translations are to be based to a form suitable for machine use. Thus in the future, if it is ever

[1] OSWALD, V. A. and FLETCHER, S. L.: *Mod. Language Forum 36,* (no 3—4) pp 1—24 (1951).

[2] BOOTH, A. D. and BOOTH, K. H. V.: “Automatic Digital Calculators” (2nd ed.) Butterworths, London, (1956).

decided to build a machine especially for translation, the material which this machine is to use will be ready. The linguistic aspect of the analysis will be described by Brandwood, but it is perhaps worth while pointing out a few of the machine techniques which are in use in our laboratory for the reduction not only of French but of any other language into English.

The first objective of translation is the looking up of words from the foreign language. The original proposals for carrying out this operation required that the unknown foreign language word should be compared successively with all the dictionary words, starting from those at the beginning of the alphabet and working through until recognition is obtained. This leads immediately to two difficulties. The first is that on the average one half of the dictionary must be searched, and the second that for inflected languages it is very unlikely that the word will be found, as it stands, in any dictionary. The hunting problem has been solved by the simple procedure (3) of binary subdivision of the dictionary by means of a programme. To make this clear, suppose that the unknown word is coded into a numerical form in which each letter is reduced to some numerical equivalent. The words in the dictionary are also coded in the form of numbers and are arranged in ascending order of numerical magnitude. The foreign language word is first subtracted from a word which is about half way between the start and finish of the dictionary. If the result of the subtraction is positive it is immediately known that the unknown word lies in the first half of the dictionary, if negative in the second half. Having performed this operation a second comparison is made in which a word either at one quarter or at three quarters of the way through the dictionary is used as the basis for comparison. By repetition of this process the dictionary is subdivided successively into one half, one quarter, one eighth, and so on and in the end the given word is located. It turns out that for a dictionary having n words an average of log_2n comparisons will be necessary, so that for a million words about 20 comparisons are required. This compares very favourably with the half million which would be necessary using the earlier technique. With APEXC this method enables any word to be located in the dictionary in a time of the order of one tenth of a second, and this is so small that it is negligible for the output speed of which the machine is capable.

The second of the two problems, that of inflected forms, is dealt with by the original technique worked out by Booth and Richens (4). In this the longest dictionary word contained in the unknown word is discovered by the subtracting process just described. This longest portion is the stem, and the dictionary is not constructed on the normal plan in which infinitives are stored, but each word is represented by its *stem*, defined as the longest segment of a word which is common to all its parts

When the stem of the unknown word has been located, the translation is in principle available. In the simplest scheme the stem translation is followed by some notes which give an account of the grammatical function of the word. Thus for example *n* for noun, *a* for adjective, *v* for verb together possibly with some indication such as *first person singular, present tense.* In the scheme which has been worked out by Brandwood however, the ending which is left when the stem has been detached is subjected to further analysis and this leads to the generation of the “ending” which must be prefixed or affixed to the English stem. The way in which this is carried out will described in Brandwood’s paper.

The next point which is of interest is the way in which idiomatic expressions can be dealt with. Idioms are of several sorts, there is the type in which words although used in normal sequences are

[3] BOOTH, A. D.: *Nature. 176,* 565 (1955).

[4] BOOTH, A. D. and LOCKE, W. N. (ed) “Machine Translation of Languages” pp. 24—46 Wiley, New York. (1955).

interpreted differently in combination than when taken singly. One of these is the French expression: *Boîte de nuit.* Which means, literally: *box of night* but the expression in French means *night-club.* The way in which such idiomatic expressions are handled by the machine is, in principle, simple. The first word, in this case *boîte,* contains in the dictionary an indication that it is not to be immediately translated but rather that subsequent words are to be examined in order to find whether these are in fact *de nuit.* In the latter event the output meaning is *night-club,* but in the event that *boîte* is not followed by the expression *de nuit,* the ordinary meaning *box* is the output. Other idiomatic expressions are of a structural type and for these the same machine analysis is involved as that used to re-order the words between the French and the English pronouns. Rearrangement is not of vital importance in translating between the languages English-French because the structure of these two languages is very similar. In German, and even more in the case of Latin and Greek, rearrangement is of vital importance if any sense is to be made of the translation at all. The calculating machine handles the rearrangement by a very simple method, each of the words in the dictionary is accompanied by a code number which indicates the structure and grammatical function of that word in the sentence. When the words of the foreign language have been presented to the machine, instead of an output being produced for each word, the machine suspends its output until it comes to a full stop or to some other point in the sentence at which it is known that the current structural unit is complete. At this point the code numbers associated with each of the words are compared with a scheme of code numbers which enables the machine to rearrange the words into the correct order. A typical example, although a very simple one, is the rearrangement of noun-adjective pairs in such a way that the adjective in the English precedes the noun where in French it may sometimes not do so. Another example lies in the complicated verb-pronoun combinations which can occur in French, and we have shown that these can be reduced to a relatively small number of coded groups and thereby rearranged into respectable English.

We are of the opinion that most of the problems which attend the translation of French into English by machine have now been solved. This does not mean that we are in a position to effect translation on a practical basis: this is not the object of our work. It remains for people who are practically interested in the subject to construct the micro-glossaries which will be needed. Before leaving the subject of French it may be perhaps worth while mentioning that the speed of translation at present obtainable on our machine is of the order of one thousand words per hour. And more depressing, the storage capacity for foreign-language words is of the order of two or three hundred. The latter restriction is due entirely to the fact that the length of a machine word corresponds only to about six alphabetical letters, since many words have a length greater than this it is necessary to take at least two machine words for each dictionary-word of the foreign language followed by at least two more machine words for the translation and analytical symbols. This means that the normal machine capacity is divided by four.

It may be mentioned in conclusion that we have been examining the application of the computing machine to language in general, French was selected as an easy starting point and one in which the author of the present paper could make contact with his linguistic advisors which would have been impossible in more exotic languages. Work is now proceeding on the analysis of the German language which in many ways poses more difficult and more interesting problems than those of French. Fortunately the invaluable analysis of Oswald and Fletcher is available and although this is not, as it stands, immediately suitable for use on our own machine it forms the basis of the analysis by which we eventually hope to produce German translations of respectable quality.