[From: *Information retrieval and machine translation*, ed. Allen Kent. International conference… September 6-12, 1959, Cleveland, Ohio; part 2 (New York: Interscience Publishers, 1961)]

CHAPTER 26

Syntactic Processing for Machine Translation and Its Implications for Cost and Speed

ANDREW D. BOOTH

Department of Numerical Automation, Birkbeck College, University of London, England

In this paper are suggested one or two tentative conclusions regarding the cost and speed of practical machine translation which arise from the results of the experiments which have been carried out at Birkbeck College during the past two or three years. To put the experiments in perspective it should be remarked that the translations which have actually been done on computing machines have been from French into English: although the problems of German to English translation have been studied extensively, this has not been programmed since none of the machines which have been hitherto available either in this laboratory or in the University of London has sufficient storage capacity to make possible even an elementary start on this language pair. For French to English translation, however, it has turned out that quite good translation can be effected in the syntactical sense with a stored programme of only about 600 words. The dictionary of actual French words, both stems and endings, with which this processing programme was associated is variable in size up to an upper limit of about 3500 words of computer storage. This means that for real language the capacity is of the order of 1000 words. It is thus quite clear that practical translation is out of the question on a machine of our sort, even if it were deemed desirable to set up such a service organization. The programmes which have been written, however, have been constructed with the idea that when a computer of of adequate storage capacity, that is between some half-million and one million words, and adequate arithmetical speed becomes available in England, the program can be "translated" very rapidly into the new computer code and used, if it seems desirable, for practical translation. In the meanwhile, by close examination of the modifications which the programme would need when it is extended to real translation on the 99% accurate scale, it has been possible to assess fairly closely the effects which such an expansion will have on its time of operation and on the storage capacities which are

As a start in this analysis it is worth enumerating a few facts about the translation process which have been devised. In the first place it

is assumed that the "partitioning" method is used for the dictionary searching technique. This means that for D entries in a complete dictionary, about $\log_2 D$ access operations are required for the location of any word. It may be argued that the bracketing method is not particularly well suited to very large scale storage devices of the "juke box" or "file drum" type or, on a more plebeian scale, to magnetic tape as an ancillary store. On the other hand, a numerical examination of the effects of using storage devices of this type on sufficiently long sequences of text has shown that the amount of look-up time involved will not differ greatly whether sorting followed by a comparison with the linear dictionary or the bracketing method applied to a random access dictionary is used.

The syntactic processing which the present French program carries out consists of the following four items: (a) the detection and removal of idioms; (b) the treatment of definite article; (c) the rearrangement of noun, adjective or adverb combinations; and (d) the processing of the complicated pronoun-verb structures which occur in French. An examination of the programs and a comparison of the number of computer words which are required to deal with the English words by which the syntactic processes are described in standard texts shows that on average one computer word is required for the replacement of one word of English descriptive text.

The experiments have used the machine M.A.C. No particular pains have been taken to optimise the machine translation programmes because in the present state of transition and development of the art the labour of optimization did not seem worth while. This being so, it is unfair to assume that the speeds which will be quoted are in any sense the best possible, even on the M.A.C. machine. It turns out that the actual operating speed of the machine on French texts is of the order 1200 words per hour. This implies, taking account of the program which is used, that the rate of execution of orders is about 50 per second. The rate 1200 words per hour takes account of input of text on punched paper tape, the production of punched paper tape by the machine and the reproduction of the punched output of a teleprinter installation. Since input and output are two machine functions which will almost certainly be quite different in a real translating machine, it is perhaps better to consider the rate of machine translation without accounting for input and output time. In this event, on M.A.C., the rate is about 2000 words per hour.

Next consider the effects of extending the program to include all of the normal French grammar. It is difficult to estimate precisely the number of computer instructions which would be required for this extensive exercise, but taking account of the present experience of one computer word for one word of English descriptive text, and inspecting a variety of French grammars, it looks as though about 60,000 words of computer programme would be required. Does this mean that the time of running of a programme of this complexity would be one hundred times slower than that of our present programme? The answer to this is quite clearly "no," since the present

programme contains branching points into which the more complicated operations, which occur far less frequently, would be inserted, These branch points are already present in the programme. They would be actuated only in the unlikely event of the curious grammatical structures with which they are designed to cope. It is not possible, of course, to assess the precise increase in time which would be involved, but independent estimates by several people in the laboratory and by visitors to whom the question has been posed suggest that the effect might be to halve the rate of translation. Those who are interested in the mechanism of such jumps which avoid complicated program moves may care to notice that there is used a standard branch instruction on the category count numbers which accompany each word in the dictionary. Thus, for example, the stem "cherch-" is accompanied by the category numbers 4 and 1020. The number 4 indicates that the stem is a verb one, the number 1020, which is substituted in the branch instruction in fact leads the programme back into its main sequence. Had it, for example, been 1021, a special program branch would be entered to deal with the contingency which does not apply to the verb "chercher." The number of these possible variants with which the present programming techniques would cope is 1024.

Having suggested that the rate of production of translation independent of input and output and of a machine of M.A.C. type will be of the order 1000 words per hour, the expense of the operation can now be estimated. In order to do useful translation a stored programme of about 60,000 words might be needed, and, if idioms are to be treated on anything like a complete scale, it is estimated that the total storage for the dictionary will be of the order of 10^6 computer words. The basic cost of a machine of M.A.C. type but with its store increased to 10^6 words in such a form that random access in a time not greater than 1/50 of a second is required, would be of the order of $50,000 for production models, although, of course, initial development costs for the large store would be somewhat greater. Assuming that such a machine would be written off in a period of about five years, and assuming very conservatively that a year consists of 3000 working hours (that is, an 8-hour shift is worked), the basic cost of the machine works out to about 3 dollars per hour. This estimate excludes maintenance and power. The power consumption of machines of M.A.C. type is negligible, and maintenance on the salary scales payable to good quality English engineering staff would be of the order of 4000 dollars per annum. Thus the total cost of translation comes out at something less than 5 dollars per thousand words.

These figures have so far excluded both input and output. It has been frequently stressed that automatic translation is unlikely to be commercially attractive so long as input texts have to be typed and checked in forms suitable for computer input. Direct input either spoken or by the recognition of printed symbols will be an essential prerequisite to the construction of the first useful translating machine. It is possible on the basis of British experience to estimate the

cost of such direct input devices, since the Solartron ERA reader is already in commercial production. Assuming the extension of facilities provided by this device from the 16 alpha-numeric symbols at present catered for to the normal range of type which would be required for translation, the cost of the device might be of the order of $60,000. Since one maintenance engineer could service both the central processor and such an input device, the addition of the device to the basic translating computer would involve only about an extra 4 dollars per hour. Thus the total cost of translation on this basis appears as something rather less than 10 dollars per thousand words. The problem of output is a less serious one. Present high speed paper tape punches can operate at speeds of up to 70,000 words per hour and are thus relatively cheap both timewise and maintenance-wise compared with the cost of the other parts of the translating unit.

It may be argued that present character reading devices make no provision for the direct input of such things as books, so that for some time to come the typist and the tape-creating equipment will be a central part of an automatic translation scheme. If this is accepted, then costs can again be computed; for example, assuming that text is typed twice for the purposes of checking and the typists are not particularly skilled in their operation, a reasonable estimate is about 1500 words per dollar. This assumes a typing rate of 50 words per minute. Thus, supposing that typists replaced the automatic reading device, the total cost of translation is rather less, in fact of the order of 6 dollars per thousand words. The advantage of the automatic recognising device is, of course, that typing rates for languages which do not employ Roman script and where the word format is not mnemonically related to that of English or to the other target language into which translation is to be made, may be far less and the error rate far higher than that which I have assumed in my estimate of 1500 words per dollar.

These estimates have been presented, based as they are upon practical experience with translation experiments, rather to give some idea of the order of magnitude of the cost which may be involved than to lay down any hard and fast predictions as to their exact numerical magnitude. It is almost certain that machines of M.A.C. type will not be used for practical translation. It is equally almost certain that the machines which are used will be less economical than a machine like M.A.C. If it were necessary to estimate what effect this might have on the ultimate cost. The conclusion would be that it is unlikely to increase this by a factor of more than about 2. Would such an increase in cost be accompanied by an increase in speed? The answer to this is more speculative, since the major time consumed in the translation process is not occupied by the programme operations themselves, but rather by access to the store, and although fast computers can increase the figure of 50 operations per second which have been mentioned by factors of anything up to 10,000, it is nevertheless true that such a reduction in the time of individual operations is accompanied by a great decrease in the available storage capacity

in the high speed sense, so that transfers to and from the high speed store would assume an increasing importance. It is not possible to estimate the effect of American programming practice on this speed, but taking into account of existing programmes written for the machine M.A.C. and for faster British machines which are currently available, it looks rather as though a speed increase by a factor of between 5 and 10 may be looked for with confidence, but that anything greater is a matter of some improbability, at least on the five year scale.