

## THRELFORD MEMORIAL LECTURE 1961

## MACHINE TRANSLATION — A CHALLENGE TO THE LINGUIST

by

Dr. A. D. BOOTH, D.Sc., Ph.D., F.Inst.P., M.Brit.I.R.E., F.I.L.(Hon.)

*Head of the Department of Numerical Automation, Birkbeck College, University of London*

SIR W. LACON THRELFORD was, in a sense, a spiritual colleague or the founder of my own Institution, Birkbeck College in the University of London. George Birkbeck was a notable character of the early Victorian days, who saw in the developing society of those days the opportunity of improving the education of industrious artisans and, as I conceive it, Sir Lacon Threlford performed a similar service for linguists. This he did by transforming The Institute of Linguists from an association of amateurs to a qualifying body of world repute in the field. I have for many years been concerned with processing languages on computing machines and it is of this field of activity that I am going to speak. Now I hope that when you decided to come to this lecture, you considered the title with some care for it shows at once a critical point in machine translation became it is ambiguous. Machine translation — a *challenge* to the linguist. I am sure that you have all seen at least the two most obvious alternative meanings of the term challenge. It could be a challenge to use the machines to the better advantage of linguistics but it could, on the other hand, be a challenge to the livelihood of linguists. I hope that by the end of this lecture you will be able to make up your own minds as to which of these particular meanings I consider to reveal the future of the subject.

Now it is customary in all lectures of this sort to give some account of the history of the subject. Histories can be long, they can be dull and they can be inaccurate. I hope that the history I am going to give will be none of these things, except in that any minuscule history of three hundred years of human endeavour must, to the extent that it cannot be all inclusive, be slightly inaccurate.

The originator of the *idea* of machine translation was the German mathematician Leibnitz. Leibnitz, in 1692, suggested firstly, that a universal mathematical machine would be possible, secondly, that such a machine could represent ideas by numbers — and those of you who are philosophers will be familiar with Leibnitz' theory of monads and its implication that words and language can be represented by binary symbols — and thirdly that such a universal machine would be capable not only of translation, which was trivial from Leibnitz' point of view, but also of original thought. Leibnitz was not a practical constructor of machines. His sole practical experiment in this field was the construction of a multiplying machine which did not work. A model of it is in the Science Museum. The reason for this "failure" was that the technology required to build such a machine did not exist.

The second pioneer of the subject was Charles Babbage, an English mathematician, philosopher, linguist and "character" who, between 1820 and 1860, invented and described in great detail the first practicable automatic computing machine. He was the father of the whole of modern computer technology. Babbage, like Leibnitz, did not produce a machine, although he contributed greatly to the progress of production engineering. His machine, even if constructed, would be too small for translation because electronic techniques had not been discovered.

The third historical figure is the Russian scientist Troyanski who, in 1933, obtained a patent on a translating machine. No complete copy of this patent specification has reached the West and it is difficult to see how Troyanski could have gone further than Leibnitz or Babbage, because the means needed for large dictionary construction were still lacking in the 1930's.

As far as *practical* experiments and *practical* proposals for translation related to a definite machine are concerned, I made the first of these in 1946. Based on small-scale surveys of linguistic data, I asserted that it was sufficient for scientific use if, instead of translating all of the words in a text, only the important technical words were translated. I developed these ideas in some detail and prepared an elementary translation programme for a computing machine. In this, the letters were represented by numbers, thus A = 01, B ... 02, C .. 03, and so on up to Z=26. In this way any word, considered as an aggregate of letters, could be represented as a number, thus : amo = 011315.

In the dictionary look-up scheme the store of the computing machine was assumed to be a rather large sheet of paper ruled with lines, each of the lines bearing a number ; line 1, line 2, line 3, line 4 and so on. The 'translation' of any word was then stored on the line whose code number was that of the foreign language concerned. Thus, on line 011315 of the Latin/English dictionary would be stored : 092712152205 = 1 (space symbol=27) love.

Now in 1946, and in fact up to about 1951, there were no machines available for doing experiments of any sort, least of all experiments in translation. Thus we were able to mature our ideas for a period of three or four years and the simple idea for using a computer store as a dictionary just described was soon found to be quite unworkable. The reason for this is that assuming for a moment that no foreign language word has more than

ten letters (the average word length in English is five letters) then the number of different words which can be constructed with ten letters is in fact  $26^{10}$  or about one hundred and forty million million !

Now linguists know that no real language has anything like one million distinct words so that even if we could build a computer with a store of this size, this process would be impracticable because most of the store would be empty, in fact 140 million blank lines, on average, would separate each line occupied by a word.

Having made this startling discovery, the immediate reaction is : dictionaries are constructed every day, and they don't have a 140 million million lines of words in them. Why not make the computer work in the same way?

To do this the computer store is arranged in the same way as a dictionary of ordinary language. Any computer storage position, or line, is divided into two parts : the first part contains the code number of the foreign language word, the second part the code number of the word or words of translation. The use of such a scheme presents certain difficulties, however, because a computer is a very stupid device whose powers of reasoning, or discrimination, extend only to doing one thing if a number held in its 'adder' is positive and another if the number is negative. Thus, on presenting the computer with an unknown word, something must be done to enable the machine to find its way through its store to the correct entry. The first way in which it was proposed to do this is a simple one. Computers work with numbers; we have represented the letter A by the code symbol 01, and so on through the alphabet. Thus the word nearest to the beginning of an ordinary dictionary is the word whose numerical code number is the smallest in the computer equivalent. In the same way the numerically greatest word in the dictionary would be called the alphabetically greatest word in an ordinary dictionary. The process of looking up a word in a dictionary is then quite simple : the computer starts at the first line in the dictionary, it compares the code number of the incoming foreign language word with the foreign language part of the entry in the dictionary. This is done by subtracting the unknown word from the dictionary entry ; if the result is negative it is clear that the required entry has not been reached. On reaching the required entry, however, the result of the subtraction becomes zero and this the machine detects, and proceeds to read out the remainder of that storage line, that is, the required translation. This process is illustrated by Fig. 1..

The process just described sounds very easy, but after one week's practical experiment it was abandoned as hopelessly uneconomical. This is because, even supposing that the dictionary contains only ten thousand words, not a large dictionary by any standards, then on average five thousand subtractions have to be made to locate the unknown word. Now, although computing machines are fast, in access to a dictionary of this size only about fifty

#### DICTIONARY SEARCH

Apres ..	0116180519	
Beau ..	0205012100	
Bleu ..	0212052100	
Chien ..	0308090514	
Dix ..	0409240000	
Gare ..	0701180500	
Ici ..	0903090000	
Il ..	0912000000	(1) —ve.
Nous ..	1415211900	
Reste ..	1805192005	(3) —ve.
Rue ..	1821050000	(4) zero.
Tard ..	2001180400	(2) +ve.
Vert ..	2205182000	
Vin ..	2209140000	
Vingt ..	2209140720	
Look up RUE	1821050000	

FIG. 1

of these comparisons would be made each second, so that in order to make five thousand comparisons, about a hundred seconds of computer time would be needed. Any normal human being could look up a word in the 100,000-word Oxford Dictionary in a time only about one tenth as great, so the machine shows a positive disadvantage.

The improved technique which we discovered is now known as the ' bracketing method'. The unknown word is compared with an entry about half-way along the dictionary. Is the foreign language code number entry greater than that of the half-way entry or less ? If greater, then we know that the word required is in the second half of the dictionary, if less, in the first half. According to the way in which the decision goes comparison is made with either the word a quarter of the way along the dictionary or three-quarters of the way. The next comparison partitions into eighths and the next to sixteenths and so on. A small amount of arithmetic shows that, for a 10,000-word dictionary, about fourteen comparisons are necessary. The operation of this process can also be seen from Fig. 1. A modern computer would take about a quarter of a second for this work, whilst a human being could not even open his dictionary in this time.

After these early ideas of using a mechanical dictionary to aid the technician in looking up words, I had the good fortune to have as a collaborator R. H. Richens of the Imperial Bureau of Plant Breeding and Genetics, Cambridge. Richens and I investigated firstly the usefulness, or otherwise, of the elementary dictionary routines, and secondly, ways in which an automatic dictionary could be made universal. One of the difficulties of the scheme just outlined is the actual technique of dictionary construction. An average dictionary containing, say, 100,000 words, does not include by any means all of the words in even a language like English. This is because languages are to some extent inflected and dictionaries do not contain all inflected forms. Thus *cherchait* does not occur in any

French dictionary, or *amat* in a Latin dictionary; instead we find that *cherchr* and *amo* are given and it is assumed that the user knows enough of the languages concerned to reduce his words to these forms. The first question which arises in mechanizing a dictionary is whether to put in all words, thereby complicating the situation, or whether to see if a better method is available. In the early days computer storage was very limited. So that although *in abstracto* all the words might be supposed to be put in the dictionary, in practice this was not possible. Richens' contribution to this field was to systematize a suggestion of mine, that inflected words might be decomposed into stems and endings. In our definition the stem is the longest portion of a word which is common to the majority of its inflected forms. The rather legal and stilted phrase, "the majority of its inflected forms," is used because irregularly-formed words may have a number of different stems and, in fact, in modern machine dictionaries we put in many complete forms of irregularly-formed words. Despite this cavil, the important point is that by this technique of stem/ending decomposition we reduce the storage capacity needed in the machine. To give a concrete example, in French technical translation, it turns out that in storing about 2,000 different everyday and technical stems and something like 200 endings to go with them, nearly one hundred thousand different inflected forms can be constructed. This principle of economy resulted from the linguistic analysis which Ronald Richens contributed when we worked together from 1947-1949. Our earliest proposals only contemplated "translations" of technical texts in the crudest sense in which only technical terms were processed and these without change in word order or insertion of even grammatical notes.

The justification for this type of approach, even without dealing with the connecting words, is shown in Examples 1, 2, 3 and 4.

Ex. 1. Passage containing only Dewey's 200 most frequent words :-

The -----, the -----, and the ----- the ----- to ----- . The ----- in the ----- of -----, and a ----- over the -----,

Ex. 2. The same passage but using 750 most frequent words :-

The -----, the -----, and the ----- turned the ----- to ----- The ----- in the ----- of light, and a ----- over the -----.

Ex. 3. Remaining words<sup>1</sup>:-

— rain ceased, — clouds parted, — — sun -----  
— muddy river — gold. — bluffs shone May - green  
— western flood -----, ----- haze hung ----- bottom lands.

The implication of these examples, for technical translation, is shown in the following passage, processed according to the original Booth - Richens technique.

Ex. 4 ' Translation ' by original Booth - Richens tech-

nique<sup>2</sup> *m*= plural, *v*= vacuous word, *z*=unspecific. Alternatives in brackets.

French :-

Il n'est pas étonnant de constater que les hormone/s de croissance ag/issent sur certain/es espèce/s, alors qu'elles sont in/opér/antes sur d'autre/s, si l'on song/e à la grand/e spécificité de ces substance/s.

English :-

*v* not is not (step) astonish *v* of establish *v* that (which) *v* hormone *m* of growth act *m* on certain species *m*, then that (which) *v* not operate *m* on of other *m* if *v* one dream (consider) *z* to *v* great *v* specificity of those substance *m*.

These techniques were codified into respectable linguistic order by Irwin Reifler, who suggested, in particular, that the "translating machine" of the future would be a composite device consisting of a machine and of human beings and that, in fact, a machine by itself would be useless. The proposal was that a machine should be accompanied by a pre-editor and a post-editor ; the pre-editor to remove all ambiguities from the incident text, the post-editor to render the output from the machine into acceptable language for the reader. The importance of this is that the pre-editor has to understand only the foreign language from which the text is drawn; he does not have to know into what language it is to be translated. The post-editor, in principle at least, need not know anything about the source language. He only has to know how to write his own language and to be a technical expert in the field concerned.

Now this is a codification of the arrangement which Richens and I had proposed, except that our post-editor was to be the scientist who was going to make use of the text when the machine had produced it. This is the picture as it existed in 1952.

I want next to make clear my own position in 1951, and to say that I have not shifted from it appreciably in the ten years which have elapsed since that time. The position essentially was this, that I envisaged translation by machine to be a time-saving aid to scientists which would enable them to get basic information regarding the content of papers written in languages which were not at that time adequately covered by translation services in this country. I refer specifically to the Russian language, in which so much scientific work has been done and in which, in the early 1950's, so little linguistic assistance was available for the scientist who wanted translations.

The next landmark in the history of machine translation was the first international conference on machine translation held at the Massachusetts Institute of Technology in the summer of 1952 ; all of the workers in the field were gathered around a small table, so you can see how limited was interest in the subject. A considerable dichotomy of ideas was revealed; some delegates were enthusiastic for the new subject, at least three delegates condemned the whole idea out of hand, and several observers from the United States Armed Forces expressed a cautious interest which has since led to the vast effort

in money and resources which the subject now commands in the United States.

In the subsequent history of machine translation it is difficult to pinpoint new ideas of real distinction. The early workers in any subject discover all of the obvious things and the rate of production of ideas decreases exponentially with time. Looking through the ten years since 1951, the following developments come to mind: Dr. John Cleave, in my own laboratory, prepared a working programme for the transcription of English to Braille by machine. The importance of this work, for machine translation, lies in the fact that, realizing the frailties and lack of capacity of available machines, Cleave decided that Braille was a useful artificial language upon which to work, and that it would be better than trying overall experiments on real languages, which were really too complicated for any machine which existed in the mid-1950's. Braille, the international language used for blind reading, was sufficiently difficult in itself because it does not consist merely of representing the letters of English by certain symbols on a Braille keyboard. There are three grades of Braille, in the first of which we do indeed have a more or less 1 - 1 correspondence of letter to symbol, in the second of which letter groups correspond to symbols, but in such a way that certain letter groups are represented by Braille symbols only if the sounds thereby produced are "euphonious", and third, a contracted 'shorthand'. Now machines are incapable of judging "euphony" in their present state, so that this subject is complicated, and in fact the way out of it which Cleave suggested, and which was largely accepted by some of the institutes for the blind, was to formalize some of the Braille rules, so that dependency on aesthetics was replaced by dependency on local context. Here we are approaching the process of translation by formalization and the use of context to resolve difficulties.

Cleave's work on Braille came to a satisfactory conclusion, and his programme and the methods proposed were described in a thesis which is in the London University Library. Are Braille transcriptions now made on machines; The answer is no, because a machine would be far more expensive than human operators for creating Braille. Furthermore, although machines themselves are far faster than a human typist, nevertheless in order to input text to a computing machine for transcription into Braille, at some stage a human typist has to transcribe the text into punching on tape. The associations for the blind use a typewriter which creates Braille directly and as this is done at the same speed as in ordinary typing, there is clearly no reason whatever to have the computing machine in between. It would produce no saving, in fact the reverse, and I ask you to remember this argument when I describe experiments in machine translation.

The next basic idea also originated in my own group, from L. Brandwood. The original 'translations' in the Booth - Richens sense involved either the outputting of

words as they stood or, at most, with some grammatical indicator. Brandwood's contribution<sup>3</sup> was to show how the stem/ending decomposition process could be accompanied by a stem-ending recomposition process in which the machine, having analysed the incident text, could then recombine the various parts of an inflected form so that, for example:  $amas=am/as=(lov-)$  (thou - est) = thou lovest, rather than, as in the Booth - Richens scheme:  $amas=am/as=love$  (2nd singular present).

After this development my group almost immediately turned to consider the subject of automatic linguistic analysis, because we had felt that the bottleneck in mechanical translation is the practical analysis of language in terms suitable for machine application and not, as has been suggested, a lack of knowledge of the theoretical principles involved. We do *not* have machine dictionaries and we do *not* have machine grammars on anything like the scale needed to make possible useful machine translation.

At about the same time as this work was going on in my laboratory, an important theoretical research was emerging at Cambridge. Here, A. F. Parker Rhodes and his colleagues were applying the techniques of lattice theory to the analysis of sentence structure and also elaborating the Thesaurus technique for the resolution of ambiguity. It is only fair to say that the practicality of these techniques, in the field of machine translation, is in dispute but, despite this, there is no question whatever that they have stimulated great activity in the field and have led to important developments in the theory of classification and data retrieval.

The last basic idea which I will mention came, in 1960, from Victor Yngve, of the Massachusetts Institute of Technology, namely the concept of 'depth'. This is based on the psychological observation that most human beings can remember only about 7 random digits, words, or ideas at any given time. From this Yngve showed that the structure of natural languages must, as far as sentence complexity is concerned, be self-limiting. He went on from this point to show how the idea could be turned inside out, so to speak, and act as an analytic tool.

One way of expressing the idea of depth is that sequences of subordinate clauses cannot be nested beyond a certain distance, because human beings would have forgotten what came before and the whole jumble would be meaningless. Incidentally, this does not apply to a computing machine. One important practical outcome of this was that Yngve was able to show how, from his depth hypothesis, to construct valid and meaningful sentences in any language given a list of words in that language and a knowledge of the linguistic grammar. At the National Physical Laboratory Conference on machine translation, held in September 1961, examples constructed by a machine on Yngve's principle were shown and the importance of the work was fairly generally accepted.

The other really important *thing*, as distinct from idea, to occur in the 1950's was that the Americans and the Russians entered the field and that they are now competing with one another for mastery. Looking at the game from the touchline, so to speak, I am filled with amusement, but I am also filled with admiration for the work in the Soviet Union, where they have realized, as no one else has, that there is a great shortage of people who understand both technology and linguistics. Furthermore, as well as realizing this difficulty, they have done something about resolving it. They have produced, over the last five years, something like 1,000 mathematical-linguists, a race which is almost non-existent elsewhere. Although the mutations caused by atmospheric radioactivity may produce a few mathematical-linguists in the West during the next twenty years, the Russians, following the principles of Lysenko, have created mathematical-linguists by the good old capitalist technique of financial inducement !

It will be appropriate here to make a few remarks on the nature of machine translation demonstrations, because accounts appear from time to time of the latest translating machine. Those of you who are as old in the ways of wickedness as I am will know all about such experiments. My own laboratory, for example, years ago had such a demonstration; we claimed to translate the French text: 'ce n'est pas un exemple de traduction par machine'. This was done with our tongue in our cheeks, and we derived great amusement from seeing this reported as an experiment in 'machine translation' in many of the newspapers.

There are two standard ways of giving a practical 'demonstration' of machine translation. In the first of these, the viewers of the experiment are presented with a pack of punched cards, on each of which is inscribed a sentence in a foreign language. They are invited to select a card which will be 'translated' by the machine. What the viewers are not told, is that punched on the card is a very good translation constructed by a very good human linguist. The second, and more recent technique, is less obvious. Viewers are given, say, a copy of *Le Temps* and invited to have the machine translate any selected paragraph into English. The method of procedure is as follows : the particular issue of *Le Temps* is translated into good English, sentence by sentence, and by a first-rate linguist. Each French sentence is then placed in the machine store and alongside it the translation into English. The passage chosen by the viewer is then presented to the machine which runs down the list already stored in its memory, finds the sentence concerned and reads the humanly constructed translation ; result : perfect 'translation'. When explained thus the method seems too obvious to be true, but if machine translation can already be done at speeds of thousands of words per hour by machine, why is it machines are not doing translation ?

An acceptable test of a machine translation claim is as follows, remembering that machines are limited in capacity so that it is not to be expected to be able to take a text at random: a list of words and available constructions is given and viewers are invited to make up their own sentences using these. If an experiment of this sort is done you can be quite sure that it is honest, if on the other hand you are told that the machine will translate any one sentence from a given mass of sentences, it should be viewed with suspicion.

In connection with the practical use of a translating machine the experiences with Braille are worth remembering. If a typist can transcribe accurately sixty thousand words of Russian text, she probably knows Russian anyway and could have dictated the translation into a recording machine. I have an intense dislike of undue publicity for a subject which is, in my estimation, still in the research field and you can judge the merit of my strictures: a) because machines are not doing practical translation at the present time, and b) because the United States Government is spending between four and ten million dollars per annum on research in the field of machine translation. This is not what you do when your machine is waiting at the door; you buy the machine, it is far cheaper !

I next propose to describe briefly the structure of an electronic computer, and then to discuss some basic ideas of mechanized linguistics. First of all the computer itself, from the point of view of translation, consists of five parts. The most important of these is the store, the electronic equivalent of a piece of notepaper, from which the machine can read, upon which it can write, and, most important, from which data can be erased. Erasure is important, not because machines are unreliable, but because electronic storage is expensive and cannot be thrown away after use. The other four parts are, respectively : the input device, basically the typist and the typewriter, the output device or printer which the machine uses to communicate the results of its operations to the outside world, the control, which is the equivalent of the human translator, and finally the arithmetical unit. The latter unit does arithmetic in normal computers, but in translation simply needs to be able to compare two numbers, and say whether one or other is bigger or smaller than the other. Technically, the modern computing machine contains a number of transistors, and has a storage device which is usually either a magnetic drum, magnetic cores, magnetic tape or all three. As an indication of speed, on some of the more modern machines arithmetic is about one million times faster than can be done by hand, so that two numbers can be compared in a hundred-thousandth of a second. For the large amounts of stored information required in linguistics, it usually takes about one-fiftieth of a second to locate a word selected at random in the dictionary if this takes the form of a magnetic drum. For very large stores, on magnetic tape, this time may increase to two or three minutes, and in this case it is

preferable to sort the text into alphabetical order before starting so that the alphabetically-stored dictionary is used in sequence. The magnetic tape may well be the principal dictionary storage technique of the future, and the National Physical Laboratory is using the ACE computing machine with a tape store in exactly this way.

The important facts about a machine are : that it is fast, that it has perfect faculties for recall of stored data, but also that the machine is infinitely stupid. The only operation of discrimination which can be assumed is that of seeing whether a number is positive, zero, or negative.

I will now discuss some of the principal ideas of mechanized linguistics. The automatic dictionary has already been mentioned and I will not say more about it. What must be remembered is that a computing machine store is a perfect and fast dictionary. It does not forget and it can locate words in times of the order of one-fiftieth of a second for a dictionary of a million words. Merely using such a dictionary can lead, in my estimation, to useful 'translations' in the fields of science and technology. On the other hand I do not believe that, in the foreseeable future, it will lead to useful translations either of literary or of scientific texts in a form suitable for publication. I shall have more to say about this when I consider the symbiosis of linguist and machine.

The second important machine application is to syntactic analysis. No translation, even at the crudest level, is likely to be generally acceptable unless account can be taken of the differences in word order between one language and another. Probably the most similar language pair is French - English, but if no attention is paid to word order change or to the differences in verb-pronoun structures, translation between these languages produces a ludicrous result. It may be said categorically that a machine can be used to overcome these difficulties. The machine can perform automatic syntactic analysis and use rearrangement procedures, 'transition rules' as we call them, so that the output text produced becomes comparatively elegant at least for texts which fall within limited subject categories such as mathematical French or some particular branch of technology.

The third of the difficulties in machine translation is that of ambiguity. On a small scale this is illustrated by such statements as: "these men are revolting," and "she cannot bear children," the meanings of which cannot be decided without the whole context as a guide. Examples can be multiplied indefinitely, but although such ambiguities appear intractable they are, in principle, resolvable on a machine. We have gone some way to finding out the technique of such resolution but the difficulty is that machine storage even at the million-word level is far too small to make such things practically possible. Considering the relative newness of the machines the situation is analogous to the human one in which you can teach Latin to an intelligent student of eleven years but it is very difficult to teach it to a baby of eleven months.

The fourth difficulty to be considered in mechanized language translation is that of idioms. Again the problem can be resolved in principle on the machine merely by storing all of the idiomatic expressions. Now linguists know that there are very many idioms, and that they change from year to year, so the same machine limitations therefore exist as with ambiguities.

Two other ideas in the theory of machine translation ought to be mentioned. The first is that of meta-language ; now meta-language is almost as capable of giving offence and creating misunderstanding as attempting to translate the word 'peace'. For example, I discovered in Moscow last year that the Russian idea of meta-language is quite different from the American one. In Russia, meta-language usually means the internal code of the machine whereas in America meta-language means a universal language of ideas which stands between all natural languages and purports to make it easier to translate from one language to another. It is difficult to illustrate this point without mathematics, but the following is a plausible argument. To translate between any pair of five languages, ABCDE, it would be possible to make dictionaries and grammars for  $A \rightarrow BCDE$ ,  $B \rightarrow ACDE$ ,  $C \rightarrow ABDE$ ,  $D \rightarrow ABCE$  and  $E \rightarrow ABCD$ , twenty such dictionaries and grammars being required. If on the other hand, we insert an intermediate, or meta-language, we merely have to translate from every language into the meta-language, and from the meta-language into each language, a total of only ten grammars and dictionaries for my example. It thus appears that there is a great saving in having a meta-language. This argument is, however, faulty since, by using, say, A as the meta-language, it is merely necessary to translate  $A \rightarrow BCDE$  and  $BCDE \rightarrow A$  — a total of only eight dictionaries and grammars. Naturally, in the present state of international amity there will be two A's — Russian and English !

Finally a few observations on the possibility of doing translation of literary quality. If asked, in discussion, if Shakespeare can be translated into Goethe, then, as a machine expert, I should say, "yes, of course — in principle". In practice, however, the limited storage capacity of the machines is again the factor which prevents realization. A literary quality translation between Shakespeare and Goethe is an experiment of the same sort as that of translating *Le Temps*, mentioned earlier. Everything that Shakespeare wrote is stored, and similarly for Goethe. To translate from one to the other, the writing of, say, Goethe, which describes roughly the same thing as Shakespeare was saying, is extracted and this is taken as the first of the Goethe sentences in the translation. Shakespeare's second sentence is then taken and the process is repeated, this time a Goethe sentence, probably from some completely different work, being extracted. When the exercise is finished, the result is what Goethe said about the situations described by Shakespeare. It could not be improved upon because it consists of Goethe's own words !

I wish now to end my lecture by returning to its subject — the challenge of the machine to the linguist. The first challenge is the following: in the field of linguistics the computing machine is in no way different from any other machine which has ever been invented to help the progress of humanity. Automation has been in existence and has been developing continuously for about five thousand years. Periodically some new feature arises, for example, the water mill, the loom, the spinning jenny, and so on. At each such advance there exist those people who say "break up the machines, the machines will destroy humanity". However, the human race has been in existence a long time and it has survived without appreciable effort. My feeling is that automation in linguistics is a thing not to be feared but, as many of my colleagues have come to realize in the last five years, to be welcomed. It will assist linguistic scholarship in many fields so that my first challenge for you is to use the machines to the best advantage and not to think of them as a means of taking away your source of livelihood.

The second challenge is directed more specifically to some of the problems encountered by linguists. There exists a very real difficulty in modern technology to translate the vast mass of material which is being published daily all over the world, especially in the Soviet Union. The trouble is that a linguist trained in the classical tradition often does not possess the special technical vocabulary which is needed to make really good and useful translations. For example, the last Threlford Memorial Lecturer, M. Delavenay, has produced an excellent book on mechanical translation.<sup>4</sup> Last year the English translation of this work appeared, produced by M. Delavenay and his wife, two expert linguists. Yet, in the English edition, many of the technical words were mistranslated in a ludicrous way ; for example, micro into milli, storage capacities altered by factors of a few thousand and so on. Now if this can be done by an author who speaks first-rate English and is an expert in his field, how much more readily can it happen with lesser mortals. The implication is clear : in the limited field of translating specific technical terms, the human linguist could well be assisted, as to vocabulary, by a machine. Each week many scientific publications appear and these tend to contain new words coined by the authors. What is the translator to do to keep abreast of this vast increase in the technical vocabulary? The answer that I would like to suggest to you is that the linguist should go into partnership with the machine, he should make use of the machine to improve his capabilities. An example of such symbiosis in a linguistic field other than translation occurred in my own College where philological scholars wished to investigate microscopical differences between variants of the text of Alfred's *Orosius*. Miss Janet Bateley has been conducting research for some years in this field and, after being persuaded to accept machine help, has admitted that the machine

accomplished in eight hours tedious work which would otherwise have occupied her for three years. The important point to be noticed here is that the machine saves human drudgery ; it does *not* replace creative human thought.

The relevance of this work to the problems of the technical translation lies in taking a new scientific text and seeing what new words the author used, and how he has used them. The machine can help because, instead of the linguist having to read sixty thousand words of text and carefully look for occurrences of new words to compare their usage in different places, the machine will very conveniently make a list of these places for human inspection. Here then is not only a challenge to the linguist, but also an aid; if you like, the pipe of peace offered by the machine designer.

The real importance of these challenges arises because there are far too few translators available. There is no unemployment among translators, quite the reverse, the waiting lists for translations are very long indeed, and my picture of the future shows the human translator and the machine as a couple existing in happy relationship. In the translation centre of the next decade the machine may be used for preliminary document searching, so that the relevant literature is first brought before the research worker in abstract form. This would result from the presentation to the machine of a set of key questions to define the field of interest, and this possibility of selection without translation has been revealed by research in machine translation on the one hand and in information retrieval on the other. It is now possible to examine a document automatically to ascertain its contents. This can be done by putting the direct question "does this week's output of Russian papers contain anything regarding X Y X W, etc.?" Alternatively the machine can be instructed to digest the papers and, using Yngve's technique of automatic sentence construction, to produce abbreviated telegraphic abstracts of the type of work described in the papers.

Having made the preliminary abstract, the second phase would be to make a rough translation if this is requested. There are two ways of doing this, the most satisfactory of which does not involve the machine, but the most practical of which does. The former method is to get a skilled linguist to sit down beside the scientist and give a rough oral run-through of the paper. The difficulty here is the shortage of skilled linguists, although there is no question that this symbiosis between two human beings is the best possible solution to the problem. The second possibility is to produce a rough machine translation. From either the quick oral run-through or the rough machine translation it can be immediately determined whether or not the paper should be translated properly. If so, the co-operation between the machine and human translator again appears to aid in the problems of special vocabulary in the manner described previously.

I have not mentioned some of the exciting symbiotic applications of machines which we have already made ; for example, in direct syntactic analysis, in chronological dating of texts, and in such heretic exercises as the establishment of the authenticity of the Pauline Corpus. These are things which can and are being done by machines which do not displace linguists, because the only people who can pose the problems are linguists. In fact they extend the activities of linguists, who, instead of working for fifty years to deal with one small section of the evidence, can now think of dealing with the whole in almost the same number of hours.

The modern world cannot produce enough skilled linguists and I do not think that anything that I have said suggests that we want fewer linguists, quite the reverse. Is the machine of any help here? As far as I know, no one has done any real work on the use of machines to teach language. But much work is in progress in the psychological field to investigate teaching machine. A machine can, for example, exercise a student in the elementary grammar and structure of a language in a way that no human teacher can do. It can keep a running account of the number of mistakes made in certain aspects of the work, and then ensure that the student exercises the weakest part of his armament most frequently. More important, it arranges that the exercises are presented in such a way that it cannot be recognized that bias is being

given to the weak links, since otherwise the student might become bored or cheat.

Finally, what of teaching the spoken language? Here, frankly, one enters the realms of science fiction, but recent work on the automatic analysis of the spoken word gives cause for hope. One of the things which a machine can do, and which is impossible for a human being, is to recognize precisely in what respect given spoken words differ from some ideal which is set. The machine can express such differences numerically and can distinguish directly the point at which the differences occur. This should form the basis of a really scientific method of training instead of the rather hit or miss ones which are now in use.

I hope that these examples have shown you that machines do form a challenge to the linguist and that this challenge is one of hope for the future rather than of the dismal competition which the timid have often predicted.

#### REFERENCES

- <sup>1</sup> Oettinger, A. G. "Automatic Language Translation." p.229. Harvard 1960.
- <sup>2</sup> Locke, W. N. and Booth, A. D. (*Ed.*). "Machine Translation of Languages." Wiley 1955.
- <sup>3</sup> Booth, A. D., Brandwood, L., and Cleave, J. P. "Mechanical Resolution of Linguistic Problems." Butterworths 1958.
- <sup>4</sup> Delavenay, E. "An Introduction to Machine Translation." Thames and Hudson 1960.