[From: *Journal of the Association for Computing Machinery*, vol.5, no.1, 1958]

Language Translation

A. F. R. BROWN

Georgetown University, Washington, D. C.

Text of paper read by the author at the meeting of the Association at Houston in June, 1957.

I must begin by admitting, as is scarcely necessary, that I am at least several months away from being able to feed a new piece of French text into a computer and have an English translation come out at the other end. In trying out my basic idea, with verbally expressed rules on filing cards, I found it was possible to arrive in about 110 hours of work at a system that would translate 220 consecutive sentences from a French chemical journal into passable English. This was during last December and January. It was so encouraging to me that it seemed reasonable to try to mechanize the system immediately and use a computer to speed up further research on the linguistic side of the problem, rather than to perfect the linguistic system by hand, so to speak, and then mechanize it. The score still remains at 220 sentences. Since February, the computer programming needed to handle the essentially linguistic part of the system has been completed, and it is now in operation on ILLIAC. It does not look French words up in a mechanical dictionary. This seems to me to be an operation whose mechanization can legitimately be left until later. Quite closely similar problems must have been solved already for many other applications of computers. At any rate, I have to convert a French sentence by hand into a series of what would be the entries for the words in my mechanical dictionary. At the other end of the process, the computer produces a translation consisting of numbers that have to be looked up in a one-for-one table of English words. This, again, seems a legitimate and indeed trivial simplification in the development stage.

Although the programming has taken so long, my expectation is that the system can be expanded and corrected indefinitely with hardly any more programming. By the end of the summer, I hope to have coded the dictionary material needed for the system to handle those 220 sentences, after which it ought to be possible to develop the system and the dictionary quite rapidly into something, that will translate most French chemical literature with, say, 90 per cent effectiveness.

My choice of French as the language to work on was obviously not dictated by a consideration of the market. The obvious choice is Russian; if I knew Russian, I too would have chosen it. However, my original intention was to demonstrate that a certain method of attack would yield results with surprising speed, and the method was one which seemed applicable to most pairs of languages.

In common with most of those who have worked on mechanical translation, I have assumed that a set of rules could be devised by which most texts on a

given subject in a given language could be transformed into texts in another language that were recognizably translations. One must also assume that the set of rules is not too large and complicated for human investigators to complete, or for a computer to apply. There are then two large questions: how to devise the rules, and how to enable the computer to apply them. In the beginning of my work I sat down to make up some rules before I had any clear idea of what sort of rules I would want; but in describing the approach now it is more convenient to begin by saying what sort of rules are involved.

To begin with, it is assumed that the French words in a sentence have been looked up in a special dictionary, and that what I shall refer to as "items" have been brought out of the dictionary, one for each French word. Each item begins with a fixed number of digits that indicate the grammatical characteristics of the French word, in fairly conventional terms. Then comes a number indicating what the French word was, and then the English equivalent that will come out as part of the translation, unless it is changed in the course of working out the sentence. After that there may follow one or more instructions, then one or more constants that are used in carrying out instructions, and finally one or more diacritics whose presence or absence may be a necessary condition for executing various instructions.

The sentence, in the form in which the computer gets it from the hypothetical dictionary search routine, contains instructions that will have to be carried out before translation is produced. These instructions correspond to the rules invented during the non-mechanical consideration of the problems. Some of the rules, however, are so general in their application that it is inefficient to plant them in individual dictionary items. For instance, there has to be a rule providing that adjectives, which mostly follow the nouns they modify in French, should be moved around to the English position, before the noun. This rule would apparently have to be included in the dictionary item for almost every French adjective. So it is more practical to have a number of general instructions, 12 of them at the moment, put at the beginning of each new sentence before instructions begin to be carried out.

After each instruction is carried out, it is discarded; and when there are no instructions left, the English words remaining in the items of the sentence are printed out; they compose, one hopes, a translation of the original French sentence. An ordinary instruction is done once and then thrown away, but a general instruction has to be done once for each item in the sentence; it is treated as though it were found at the right moment in each item in turn.

The order in which instructions are to be carried out has to be controlled very carefully, to avoid conflict. The most important reason for this is the need to make all the decisions that depend on French word order before the items are shuffled around into English word order. The sequence of execution is fixed by beginning each instruction with a priority number, within a somewhat arbitrary range of one to 126. Whenever the computer has to decide which instruction to follow next, it looks for the one with the lowest priority number. In case of a tie, the instruction occurring earlier in the sentence is done first. Within each

item, the instructions are listed in the order in which they are to be done, so that actually the computer only has to look at the first of the remaining instructions in each item, and at the first remaining one in the series of general instructions, to decide which one to take up next.

Besides its priority number, seven binary digits in the present system, an instruction contains the name of an operation, nine digits, and a parameter, three digits. The parameter has a rather minor function, enabling reference to be made to comparison constants included in the same item with the instruction. To carry out the instruction, the name is used to refer to the operation, a series of consecutive computer words carried permanently in magnetic drum storage. The operation may begin with a series of one or more comparison constants which can be referred to by number, and after any such constants follows a series of words that might be called sub-instructions. These are converted by an interpretive routine into a program of sub-operations, which are computer routines permanently stored in the Williams memory. Such a program contains orders for making changes in the sentence when appropriate. It may contain logical decisions and loops of all kinds, but unlike a computer program it does not have much ability to alter itself.

There are 48 basic sub-operations, and all the operations I have concocted or imagined so far can be conveniently programmed in terms of them by reference to simple tables, without having to do any computer programming. The various sub-operations can make the item containing the current instruction the "current" item, to be looked at or altered or moved, or can make the item before or after the presently current one into the new current one. They can ask whether the current item has the grammatical characteristics, or the English word, or the French word, indicated by one of the comparison constants; or whether it contains a given instruction or diacritic. Or they can look forward or backward in the sentence until they find an item that satisfies some such condition, making it the new current item if it is found, and returning a negative answer if none is found. Other sub-operations can change the grammatical characteristics or the English of an item, or insert an instruction or a diacritic into it, or delete or insert a whole item, or change the order of items in the sentence.

The whole process of translation by this method can be described as a double interpretive routine. Raw material is got from the dictionary and then subjected to the first interpretive routine, which causes instructions to be performed in the correct order, and a series of English words to be printed out after the last instruction is performed. Each instruction, in turn, takes an operation from storage and interprets it as a program of sub-operations. Now the sub-operations, and the interpretive routines, are so general as to be almost independent of what languages are concerned in the translation. So the method has the possible advantage that one master program, with only slight changes, might be used for several different sorts of translation. A different dictionary would have to be used in each case, of course, and a different set of instructions would have to be stored on the drum. But, as the basic program has taken several months of part-

time work to get ready, it is encouraging to think that this work may not have to be repeated if I should get mechanical translation of chemical French into actual operation, and then turn to some language more in demand at the present time.

It remains to describe the method by which the rules, and the dictionary items for them to work on, have been arrived at. I opened a recent French chemical journal at random, went to the beginning of the article, and set out to formulate verbal rules that would translate the first sentence. It had about forty words, and it took ten hours to work out the rules. Turning to the second sentence, I added new items to the dictionary, invented new rules, and modified existing rules until the system would handle both sentences. The third sentence was attacked in the same way, and so on up to 220.

The time required to add each new sentence to the repertory tapered off rapidly, and by the two hundredth sentence it averaged about fifteen minutes per sentence. And this time was mostly consumed in shuffling cards in and out of the file, and writing cards for new items of vocabulary, without much thought necessary. To my own satisfaction at least, this showed that by the time two hundred sentences of running text have been processed in this way, in French at least, most of the major difficulties have been met and solved moderately well. Further progress, once the mechanical version of the system is working smoothly, should be very rapid Fresh text can be fed into the machine in batches of say ten sentences at a time. If a mechanical dictionary system is already working, then something purporting to be a translation will be produced for each sentence, and the existence of new French words that need to have dictionary items written will be signalled in the translations. If dictionary lookup is still a hand operation, then the new dictionary items have to be written, by analogy with items for comparable words, before the fresh text is fed in. In either case, some of the sentences will be translated acceptably, and others will not. The unacceptable translations will show fairly clearly where the existing system goes wrong. Some operations will have to be modified, and new instructions will have to be added. These can be tested without much trouble on a selection of the earlier sentences, to make sure that the new rules are not conflicting with old ones, and that the old ones have not been spoiled in modification.

Ultimately, a point should be reached at which 90 percent or 95 percent of the sentences in each new batch of text are adequately translated, with no prior additions to the dictionary needed, and it might then be claimed that a useful, though no doubt uneconomic, system for machine translation of chemical French had been achieved. The objection may be made that the day of 90 percent effectiveness will be a long time away if it is approached simply by going ahead from one sentence of text to the next, and accumulating the system accordingly. I do not think this objection is valid. In the first place, every little rule that is added to the system as I am trying to build it up would have to be discovered and formulated in some form or other in any method of research. The job might be done in large batches instead of bit by bit, yet the size of the job could not be much different. In the second place, the sentence by sentence approach is

the only one I can see in which a computer can he made to do most of the dirty work, leaving the investigator to develop the system according to the necessities indicated by the computer.

The opposite approach to the one I am following involves the attempt to find a few very powerful rules, rather than a lot of rules with limited application. The attempt is often to make rules that will recognize large syntactic patterns and units in the input language, produce the corresponding patterns and units in the output language, and then, so to speak, fill in the blanks with the correct words of the output language. This may turn out to be the best method for converting, say, Japanese sentences into English ones, since the patterns are so different. But if one takes a sentence in a European language and translates it into English, one generally sees that a few special idiom rules, and some well-defined changes of word order, would convert the word-for-word translation into an acceptable translation. And the easiest way to provide for these is to have a rule for translating each idiom attached to one of the words in the idiom, and to make up specific rules for the changes of word order. No one has even hinted, so far, at a description of French linguistic structure that does not involve interlocking structures within structures. So I doubt that a "technological breakthrough" is likely by which the rules could be made not only few in number and powerful, but also simple enough to represent much of a net gain.

The method of a few powerful rules makes research much more difficult. In the first place, rules of this kind will take much time and thought to devise, and at least the rough drafts of them have to be made up before the system can be tried out on any text. While if one progresses from one sentence to the next, making up reasonable but ad hoc rules as one goes along, one always has a system that will actually handle all the text that has been processed so far, and will point out its own specific inadequacies as more text is processed. In the second place, it is hard to tinker with a system that consists mainly of powerful rules. A powerful rule has to be so involved that any modification may well mean rewriting it from scratch, and this may mean fresh computer programming. A system of many small rules, however, can be modified at one point, by changing or inventing one rule, without necessarily disturbing the rest. And as new operations can be composed of standard sub-operations, no new programming is needed.

Another method of research which I have not been tempted to use involves making a number of studies of general problems, and then combining the results. One may study prepositions in general, and plausibly solve the problems connected with translating them. Separately, one may solve the problems of pronouns, and of verb tenses and moods, and so on. But before these solutions can be of actual use, they have to be combined into one over-all system, and it would be extremely difficult to know whether the tactics used on the pronouns, say, might not be disturbing the evidence on which the treatment of the past participle was to be based. All this is avoided if the system is developed as a whole from the very beginning, all its parts being made to fit each other by controlling the order in which instructions are carried out.

A further disadvantage of making general studies is the temptation to study too much and try to solve too many problems. To prevent a waste of effort, supposing French chemical text is being considered, a big sample of that sort of text has to be studied to see how many of the resources of the French language in general are used often enough to bother with. Here again, an investigator who builds up a total system by working through continuous text is prevented from wasting time on too many of the complexities which he can imagine, but which are in practice very uncommon.

There are three specific points I would like to discuss next, which are brought up almost too often in discussions of machine translation. My excuse is that I think it can be shown that the problems are not nearly as difficult as they are generally made out to be. First, idioms. For example, it was some time before I realized that *eau oxygenée* was not to be literally translated *oxygenated water*, but meant *hydrogen peroxide* in English. A simple rule now provides for this; to make sure it is brought into play as rarely as possible, it is planted in the item for the least common word in the idiom. Thus the item for *oxygenée* contains an instruction which is eventually interpreted: "Look at the item next preceding. Does it contain the French word *eau?* If so, change its English word to *hydrogen peroxide,* and delete this item (the one for *oxygenée).*" Most idioms can be provided for by equally straightforward instructions, and in fact they seem to be the least difficult of the problems of mechanical translation.

The second question is that of how to store vocabulary in a language with many inflections; whether there should be one item in the dictionary for each form of a verb, for example, or whether there should be one item for the stem of each regular verb, with the affixes listed separately. The assumption too generally made is that having a separate item for every inflected form will make the glossary too large to be practical. This may be true, but in the first place nobody knows, or at any rate nobody has yet stated, how many binary digits will compose the average dictionary item in his system. For my own system, I would estimate about 120 digits, plus five for each letter of the English word or words contained in the item; say 160 bits on an average. In any case, where no such estimate has been made, it is fruitless to worry about the size of one's dictionary. In the second place, it will probably be a couple of years before mechanical translation gets into any sort of commercial production, and by that time, we are presumably confident, large advances will be made in the techniques of data storage. It does not seem ridiculous to hope that storing and referring to an immensely large dictionary will be quite practicable. In the meantime, the most sensible strategy would seem to be to develop a system using separable endings, but in such a way that very few changes need to be made if a dictionary containing each individual form of an inflected word turns out to be practical. In my own work, I think I have achieved a method which is compatible in this way, by listing the endings that each stem can take immediately after that stem in the dictionary, each one followed by the indications of its fraction of the total meaning. This might seem wasteful of space, compared with listing every suffix just once in a master-list of suffixes. But it has the advantage of compatibility with a

system of completely multiple storage, and it greatly simplifies the use of the dictionary in certain other ways.

The third often-raised point is how to refer to a dictionary rapidly, if it has to be carried on a one-dimensional medium like magnetic tape or punched tape. This, again, is not as awkward as it may seem. On the ILLIAC, for instance, it looks as though 200,000 binary digits of drum storage will be available for use in dictionary lookup; this amount of space will probably hold the items for six to eight hundred different French words comfortably. A continuous passage of up to 1000 words of text may be handled with this number of different items. So if a passage of that length is first read, the words can be, in effect, sorted into alphabetical order; and one reading of the dictionary tape or tapes from end to end can be used for extracting all the necessary items. For the ILLIAC, the sheer bulk of the paper tape may well make this impractical, but if the entire dictionary could be got onto a single tape which the ILLIAC could read through as a single operation, and if the delay for rewinding the dictionary tape after each use could be eliminated, the lookup time might be of the order of five seconds per item. This is not an impressive speed, but the ILLIAC was designed with quite different uses in mind. The same system on an advanced IBM computer could probably consult the dictionary at the rate of five or ten items per second.

This method can perhaps be pushed a stage further. If there were a hundred thousand items in the dictionary, and enough drum storage for ten thousand French words, without their items, a stretch of about eight thousand words might be read, containing say five thousand different words. A temporary dictionary tape of five thousand items might then be selectively copied from the complete dictionary tape; and this smaller tape could be consulted five times in translating the eight-thousand-word piece of text. The whole process would be rather faster than consulting the master dictionary tape five times in the first place. However, this process looks so involved that I keep a sneaking hope that some more accessible storage medium than tape will be available soon for very large-scale storage[1].

Actually, the size of the dictionary needed for translating chemical literature from French into English will not be nearly so large as at first I supposed. The great bulk of the inorganic terminology does not have to be entered in the dictionary at all. Suppose the dictionary does contain entries for the singular and plural forms of *sulfurique.* Then it is not necessary to provide entries for *nitrique.* One simply provides a rule that when a word with the suffix *-ique* or *-iques* is not found in the dictionary, it is to be given an item containing an English word made by changing the ending to *-ic,* and otherwise identical with the dictionary item for *sulfurique(s).* So large families of nouns and adjectives can be disposed of by listing only one member of each in the dictionary. Organic terminology, though it contains a much vaster number of words, should be even easier to provide for. The names of organic compounds are almost all recognizable as such by their suffixes. So instead of listing them in the dictionary, one provides a

[1] See the paper by D. M. BAUMANN in this issue.—*Ed.*

rule that any French word not found in the dictionary but possessing an organic ending is to be given an item consisting of an English word spelt just like the French one—let the accents fall where they may—with the instructions and diacritics appropriate to all names of organic compounds, whatever they turn out to be. Occasionally, of course, an odd but still perfectly recognizable spelling would appear in the translation.

The same plan could probably be adapted for translation among most European languages. It would not be available, of course, for translation between European and Asiatic languages. Even Russian, from the little I know of it, would give the machine translator a hard time with its chemical vocabulary, though I think some of the dictionary storage could be saved by systematization of this kind. At any rate, I feel lucky to have picked French to try my hand on for the present.