1 • Translation*

WARREN WEAVER

There is no need to do more than mention the obvious fact that a multiplicity of languages impedes cultural interchange between the peoples of the earth, and is a serious deterrent to international understanding. The present memorandum, assuming the validity and importance of this fact, contains some comments and suggestions bearing on the possibility of contributing at least something to the solution of the world-wide translation problem through the use of electronic computers of great capacity, flexibility, and speed.

The suggestions of this memorandum will surely be incomplete and naïve, and may well be patently silly to an expert in the field—for the author is certainly not such.

A War Anecdote—Language Invariants

During the war a distinguished mathematician whom we will call *P*, an ex-German who had spent some time at the University of Istanbul and had learned Turkish there, told W. W. the following story.

A mathematical colleague, knowing that *P* had an amateur interest in cryptography, came to *P* one morning, stated that he had worked out a deciphering technique, and asked *P* to cook up some coded message on which he might try his scheme. *P* wrote out in Turkish a message containing about 100 words; simplified it by replacing the Turkish

* *Editors' Note:* This is the memorandum written by Warren Weaver on July 15, 1949. It is reprinted by his permission because it is a historical document for machine translation. When he sent it to some 200 of his acquaintances in various fields, it was literally the first suggestion that most had ever seen that language translation by computer techniques might be possible.

letters ç, ğ, ĭ, ö, ş, and ü by c, g, i, o, s, and u respectively; and then, using something more complicated than a simple substitution cipher, reduced the message to a column of five-digit numbers. The next day (and the time required is significant) the colleague brought his result back, and remarked that they had apparently not met with success. But the sequence of letters he reported, when properly broken up into words, and when mildly corrected (not enough correction being required really to bother anyone who knew the language well), turned out to be the original message in Turkish.

The most important point, at least for present purposes, is that the decoding was done by someone who did not know Turkish, and did not know that the message was in Turkish. One remembers, by contrast, the well-known instance in World War I when it took our cryptographic forces weeks or months to determine that a captured message was coded from Japanese; and then took them a relatively short time to decipher it, once they knew what the language was.

During the war, when the whole field of cryptography was so secret, it did not seem discreet to inquire concerning details of this story; but one could hardly avoid guessing that this process made use of frequencies of letters, letter combinations, intervals between letters and letter combinations, letter patterns, etc., *which are to some significant degree independent of the language used.* This at once leads one to suppose that, in the manifold instances in which man has invented and developed languages, there are certain invariant properties which are, again not precisely but to some statistically useful degree, common to all languages.

This may be, for all I know, a famous theorem of philology. Indeed the well-known *bow-wow, woof-woof,* etc. theories of Müller and others, for the origin of languages, would of course lead one to expect common features in all languages, due to their essentially similar mechanism of development. And, in any event, there are obvious reasons which make the supposition a likely one. All languages—at least all the ones under consideration here—were invented and developed by *men;* and all men, whether Bantu or Greek, Islandic or Peruvian, have essentially the same equipment to bring to bear on this problem. They have vocal organs capable of producing about the same set of sounds (with minor exceptions, such as the glottal click of the African native). Their brains are of the same general order of potential complexity. The elementary demands for language must have emerged in closely similar ways in different places and perhaps at different times. One would expect wide superficial differences; but

it seems very reasonable to expect that certain basic, and probably very nonobvious, aspects be common to all the developments. It is just a little like observing that trees differ very widely in many characteristics, and yet there are basic common characteristics—certain essential qualities of "tree-ness,"—that all trees share, whether they grow in Poland, or Ceylon, or Colombia. Furthermore (and this is the important point), a South American has, in general, no difficulty in recognizing that a Norwegian tree *is* a tree.

The idea of basic common elements in all languages later received support from a remark which the mathematician and logician Reichenbach made to W. W. Reichenbach also spent some time in Istanbul, and, like many of the German scholars who went there, he was perplexed and irritated by the Turkish language. The grammar of that language seemed to him so grotesque that eventually he was stimulated to study its logical structure. This, in turn, led him to become interested in the logical structure of the grammar of several other languages; and, quite unaware of W. W.'s interest in the subject, Reichenbach remarked, "I was amazed to discover that, for (apparently) widely varying languages, the basic logical structures have important common features." Reichenbach said he was publishing this, and would send the material to W. W.; but nothing has ever appeared.

One suspects that there is a great deal of evidence for this general viewpoint—at least bits of evidence appear spontaneously even to one who does not see the relevant literature. For example, a note in *Science,* about the research in comparative semantics of Erwin Reifler of the University of Washington, states that "the Chinese words for 'to shoot' and 'to dismiss' show a remarkable phonological and graphic agreement." This all seems very strange until one thinks of the two meanings of "to fire" in English. Is this only happenstance? How widespread are such correlations?

Translation and Computers

Having had considerable exposure to computer design problems during the war, and being aware of the speed, capacity, and logical flexibility possible in modern electronic computers, it was very natural for W. W. to think, several years ago, of the possibility that such computers be used for translation. On March 4, 1947, after having turned this idea over for a couple of years, W. W. wrote to Professor Norbert Wiener of Massachusetts Institute of Technology as follows:

One thing I wanted to ask you about is this. A most serious problem, for UNESCO and for the constructive and peaceful future of the planet, is the problem of translation, as it unavoidably affects the communication between peoples. Huxley has recently told me that they are appalled by the magnitude and the importance of the translation job.

Recognizing fully, even though necessarily vaguely, the semantic difficulties because of multiple meanings, etc., I have wondered if it were unthinkable to design a computer which would translate. Even if it would translate only scientific material (where the semantic difficulties are very notably less), and even if it did produce an inelegant (but intelligible) result, it would seem to me worth while.

Also knowing nothing official about, but having guessed and inferred considerable about, powerful new mechanized methods in cryptography—methods which I believe succeed even when one does not know what language has been coded—one naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say: "This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode."

Have you ever thought about this? As a linguist and expert on computers, do you think it is worth thinking about?

Professor Wiener, in a letter dated April 30, 1947, said in reply:

Second—as to the problem of mechanical translation, I frankly am afraid the boundaries of words in different languages are too vague and the emotional and international connotations are too extensive to make any quasimechanical translation scheme very hopeful. I will admit that basic English seems to indicate that we can go further than we have generally done in the mechanization of speech, but you must remember that in certain respects basic English is the reverse of mechanical and throws upon such words as *get* a burden which is much greater than most words carry in conventional English. At the present time, the mechanization of language, beyond such a stage as the design of photoelectric reading opportunities for the blind, seems very premature. . . .

To this, W. W. replied on May 9, 1947:

I am disappointed but not surprised by your comments on the translation problem. The difficulty you mention concerning Basic seems to me to have a rather easy answer. It is, of course, true that Basic puts multiple use on an action verb such as *get*. But, even so, the two-word combinations such as *get up, get over, get back,* etc., are, in Basic, not really very numerous. Suppose we take a vocabulary of 2,000 words, and admit for good measure all the two-word combinations as if they were single words. The vocabulary is still only four million: and that is not so formidable a number to a modern computer, is it?

Thus this attempt to interest Wiener, who seemed so ideally equipped to consider the problem, failed to produce any real result. This must in fact be accepted as exceedingly discouraging, for, if there

are any real possibilities, one would expect Wiener to be just the person to develop them.

The idea has, however, been seriously considered elsewhere. The first instance known to W. W., subsequent to his own notion about it, was described in a memorandum dated February 12, 1948, written by Dr. Andrew D. Booth who, in Professor J. D. Bernal's department in Birkbeck College, University of London, had been active in computer design and construction. Dr. Booth said:

A concluding example, of possible application of the electronic computer, is that of translating from one language into another. We have considered this problem in some detail, and it transpires that a machine of the type envisaged could perform this function without any modification in its design.

On May 25, 1948, W. W. visited Dr. Booth in his computer laboratory at Welwyn, London, and learned that Dr. Richens, Assistant Director of the Bureau of Plant Breeding and Genetics, and much concerned with the abstracting problem, had been interested with Dr. Booth in the translation problem. They had, at least at that time, not been concerned with the problem of multiple meaning, word order, idiom, etc., but only with the problem of mechanizing a dictionary. Their proposal then was that one first "sense" the letters of a word, and have the machine see whether or not its memory contains precisely the word in question. If so, the machine simply produces the translation (which is the rub; of course "the" translation doesn't exist) of this word. If this exact word is not contained in the memory, then the machine discards the last letter of the word, and tries over. If this fails, it discards another letter, and tries again. After it has found the largest initial combination of letters which *is* in the dictionary, it "looks up" the whole discarded portion in a special "grammatical annex" of the dictionary. Thus confronted by *running,* it might find *run* and then find out what the ending *(n)ing* does to *run.*

Thus their interest was, at least at that time, confined to the problem of the mechanization of a dictionary which in a reasonably efficient way would handle *all forms* of all words. W. W. has no more recent news of this affair.

Very recently the newspapers have carried stories of the use of one of the California computers as a translator. The published reports do not indicate much more than a word-into-word sort of translation, and there has been no indication, at least that W. W. has seen, of the proposed manner of handling the problems of multiple meaning, context, word order, etc.

This last-named attempt, or planned attempt, has already drawn forth inevitable scorn, Mr. Max Zeldner, in a letter to the *Herald Tribune* on June 13, 1949, stating that the most you could expect of a machine translation of the fifty-five Hebrew words which form the 23d Psalm would start out **Lord my shepherd no I will lack,** and would close **But good and kindness he will chase me all days of my life; and I shall rest in the house of Lord to length days.** Mr. Zeldner points out that a great Hebrew poet once said that translation "is like kissing your sweetheart through a veil."

It is, in fact, amply clear that a translation procedure that does little more than handle a one-to-one correspondence of words cannot hope to be useful for problems of *literary* translation, in which style is important, and in which the problems of idiom, multiple meanings, etc., are frequent.

Even this very restricted type of translation may, however, very well have important use. Large volumes of technical material might, for example, be usefully, even if not at all elegantly, handled this way. Technical writing is unfortunately not always straightforward and simple in style; but at least the problem of multiple meaning is enormously simpler. In mathematics, to take what is probably the easiest example, one can very nearly say that each word, within the general context of a mathematical article, has one and only one meaning.

The Future of Computer Translation

The foregoing remarks about computer translation schemes which have been reported do not, however, seem to W. W. to give an appropriately hopeful indication of what the future possibilities may be. Those possibilities should doubtless be indicated by persons who have special knowledge of languages and of their comparative anatomy. But again, at the risk of being foolishly naïve, it seems interesting to indicate four types of attack, on levels of increasing sophistication.

Meaning and Context

First, let us think of a way in which the problem of multiple meaning can, in principle at least, be solved. If one examines the words in a book, one at a time as through an opaque mask with a hole in it one word wide, then it is obviously impossible to determine, one at a time, the meaning of the words. "Fast" may mean "rapid"; or it may mean "motionless"; and there is no way of telling which.

But, if one lengthens the slit in the opaque mask, until one can see not only the central word in question but also say N words on either side, then, if N is large enough one can unambiguously decide the meaning of the central word. The formal truth of this statement becomes clear when one mentions that the middle word of a whole article or a whole book is unambiguous if one has read the whole article or book, providing of course that the article or book is sufficiently well written to communicate at all.

The practical question is: "What minimum value of N will, at least in a tolerable fraction of cases, lead to the correct choice of meaning for the central word?

This is a question concerning the statistical semantic character of language which could certainly be answered, at least in some interesting and perhaps in a useful way. Clearly N varies with the type of writing in question. It may be zero for an article known to be about a specific mathematical subject. It may be very low for chemistry, physics, engineering, etc. If N were equal to 5, and the article or book in question were on some sociological subject, would there be a probability of 0.95 that the choice of meaning would be correct 98% of the time? Doubtless not: but a statement of this sort could be made, and values of N could be determined that would meet given demands.

Ambiguity, moreover, attaches primarily to nouns, verbs, and adjectives; and actually (at least so I suppose) to relatively few nouns, verbs, and adjectives. Here again is a good subject for study concerning the statistical semantic character of languages. But one can imagine using a value of N that varies from word to word, is zero for *he, the,* etc., and needs to be large only rather occasionally. Or would it determine unique meaning in a satisfactory fraction of cases, to examine not the $2N$ adjacent *words,* but perhaps the $2N$ adjacent *nouns?* What choice of adjacent words maximizes the probability of correct choice of meaning, and at the same time leads to a small value of N?

Thus one is led to the concept of a translation process in which, in determining meaning for a word, account is taken of the immediate ($2N$ word) context. It would hardly be practical to do this by means of a generalized dictionary which contains all possible phases $2N + 1$ words long: for the number of such phases is horrifying, even to a modern electronic computer. But it does seem likely that some reasonable way could be found of using the micro context to settle the difficult cases of ambiguity.

Language and Logic

A more general basis for hoping that a computer could be designed which would cope with a useful part of the problem of translation is to be found in a theorem which was proved in 1943 by McCulloch and Pitts.[1] This theorem states that a robot (or a computer) constructed with regenerative loops of a certain formal character is capable of deducing any legitimate conclusion from a finite set of premises.

Now there are surely alogical elements in language (intuitive sense of style, emotional content, etc.) so that again one must be pessimistic about the problem of *literary* translation. But, insofar as written language is an expression of logical character, this theorem assures one that the problem is at least formally solvable.

Translation and Cryptography

Claude Shannon, of the Bell Telephone Laboratories, has recently published some remarkable work in the mathematical theory of communication.[2] This work all roots back to the statistical characteristics of the communication process. And it is at so basic a level of generality that it is not surprising that his theory includes the whole field of cryptography. During the war Shannon wrote a most important analysis of the whole cryptographic problem, and this work is, W. W. believes, also to appear soon, it having been declassified.

Probably only Shannon himself, at this stage, can be a good judge of the possibilities in this direction; but, as was expressed in W. W.'s original letter to Wiener, it is very tempting to say that a book written in Chinese is simply a book written in English which was coded into the "Chinese code." If we have useful methods for solving almost any cryptographic problem, may it not be that with proper interpretation we already have useful methods for translation?

This approach brings into the foreground an aspect of the matter that probably is absolutely basic—namely, the statistical character of the problem. "Perfect" translation is almost surely unattainable. Processes, which at stated confidence levels will produce a translation which contains only X per cent "error," are almost surely attainable.

And it is one of the chief purposes of this memorandum to emphasize that *statistical semantic* studies should be undertaken, as a necessary preliminary step.

The cryptographic-translation idea leads very naturally to, and is in fact a special case of, the fourth and most general suggestion: namely, that translation make deep use of language invariants.

Language and Invariants

Indeed, what seems to W. W. to be the most promising approach of all is one based on the ideas expressed on pages 16–17—that is to say, an approach that goes so deeply into the structure of languages as to come down to the level where they exhibit common traits.

Think, by analogy, of individuals living in a series of tall closed towers, all erected over a common foundation. When they try to communicate with one another, they shout back and forth, each from his own closed tower. It is difficult to make the sound penetrate even the nearest towers, and communication proceeds very poorly indeed. But, when an individual goes down his tower, he finds himself in a great open basement, common to all the towers. Here he establishes easy and useful communication with the persons who have also descended from their towers.

Thus may it be true that the way to translate from Chinese to Arabic, or from Russian to Portuguese, is not to attempt the direct route, shouting from tower to tower. Perhaps the way is to descend, from each language, down to the common base of human communication—the real but as yet undiscovered universal language—and then re-emerge by whatever particular route is convenient.

Such a program involves a presumably tremendous amount of work in the logical structure of languages before one would be ready for any mechanization. This must be very closely related to what Ogden and Richards have already done for English—and perhaps for French and Chinese. But it is along such general lines that it seems likely that the problem of translation can be attacked successfully. Such a program has the advantage that, whether or not it lead to a useful mechanization of the translation problem, it could not fail to shed much useful light on the general problem of communication.

REFERENCES

1. Warren S. McCulloch and Walter Pitts, *Bull. math. Biophys.*, no 5, pp. 115–133, 1943.

2. For a very simplified version, see "The Mathematics of Communication," by Warren Weaver, *Sci. Amer.*, vol. 181, no. 1, pp. 11–15, July 1949. Shannon's original papers, as published in the *Bell Syst. tech. J.*, and a longer and more detailed interpretation by W. W. are about to appear as a memoir on communication, published by the University of Illinois Press. A book by Shannon on this subject is also to appear soon. [A joint book, *The Mathematical Theory of Communication,* by Shannon and Weaver, was published by the University of Illinois Press in 1949—*Editors' Note*]