[Reprint from The University of Texas *Studies in Literature and Language*
Volume III, No. 2, Summer 1961]

# MACHINE TRANSLATION: AGENT OF THE HUMANITIES

*By* W. P. LEHMANN

OFTEN WHEN A NEW AREA OF INVESTIGATION is developed it is given a label which conceals its possibilities from those not involved. *Machine translation* is a misleading label, but after its general adoption there is little point in attempting to replace it or the alternate label, *mechanical translation,* even though, to some, "machine" suggests a little black box, to the more sophisticated, one of the costly new instruments with flashing lights, often preceded by the initials *I. B.* Either concept will be inadequate when one of the specialists in a complex new study of human communication uses *machine* as a virtual equivalent of *theory.* His use of the word *translation* may correspond more closely to its general definition, yet he would scarcely admit that his only interest consists in providing his colleagues in science row with accurate, though dull, versions of publications outside their native language. These disclaimers about the practical role of machine translation may seem baffling to other scholars in the humanities who for the first time have seen large amounts of money entrusted to some of their fellows, eagerly, without the patronizing smile with which dispensers of research funds often award several hundred dollars for a project in the humanities. To understand what machine translation is, what its aims and possibilities are, we must review the accomplishments of linguistics, and the problems it has not solved.

If we define *linguistics* as the study of language, even if we sharpen our definition to the systematic study of language, we must acknowledge it as one of man's oldest areas of organized research. The Akkadians studied Sumerian systematically in the third millenium before Christ, for it was of religious and economic importance to them. Similar aims subsequently prompted the study of language in other parts of the world, of Sanskrit in India, of Latin in Europe, even of modern languages in Europe and America during the past few centuries. Yet linguistics as undertaken in these various areas never developed very far. Herodotus probably accomplished more in anthropology than any Greek in linguistics. And although Panini's work on grammar was a revelation to Europeans at the beginning of the nineteenth century, the Indian Kamasutra could teach the disciples of Kinsey more than Panini's sutras can contemporary linguists. Linguistics has made greater progress in the last four decades than in all the previous millennia of work on language. Yet in the last decade much linguistic activity has

been devoted to sterile questions: many linguists have concerned themselves with such problems as whether the entities before the vowels in *church* and *judge* are unitary or not; others have based analyses on introspection, positing units without consulting available mechanical means of analysis or at best using such means of analysis amateurishly. As a result some observers have begun to speak of an era of decline, of an Alexandrian period in linguistics. It would be remarkable if the study of such a complex phenomenon as language flourished brightly for several decades and then entered a period of doldrums. The reason lies partly in the methodology applied in linguistics, partly in the complexity of language. Since the latter is more readily dealt with, we will first examine how it stifled the development of linguistics.

The brilliant development of linguistics in the twenties and thirties of this century derived from the understanding that language must be described in terms of its own structure. Just as chemists analyze the entities with which they deal for elements rather than for their relation to the outside world, such as their utility, so linguists came to define linguistic entities in terms of their behavior within language rather than of their meaning, i.e., their relation to the outside world. For a linguist a noun is not the name of a person, place, or thing, but rather a class of elements that shares some common behavior, such as the possibility of making a plural by the addition of a suffix (*dog* : *dogs*) or of standing after certain other entities (*the dog*, but not *the went*). Examining a language to determine in this way its classes is a complex affair. It can be done fairly readily for sounds and forms, though even here there are problems; Haugen, for example, pointed out[1] that it would take considerable study of English to discover the final [mb] group which exists in English only in *iamb* and *dithyramb*, for some speakers. Apart from such rarities, English has been well analyzed in phonology and morphology. But these are the least complex levels of language. The chief function of language is to communicate, and for this the phonological and morphological levels scarcely proceed beyond primitive situations. Even to determine the sound system and the form system of any language is not simple. To proceed further towards a description of the syntactic system and the semantic system, even intuitively, is a task on which most of us struggle for a lifetime, with our native language. For a description of the syntactic system of English we must know why a foreigner is wrong when he says "She gave me the informations" and "I am believing to go" on the pattern of "She gave me the directions" and "I am planning to go." No grammar or dictionary exists

[1] *Language*, XXVII (1951), 291.

which contains such a description. Yet without it we have failed to describe English completely. We may of course adopt a puerile escape route and limit linguistics to phonology and morphology. Such a restricted subject would be of approximately the same social use as classification of sea shells. Under whatever label it is accomplished, a complete analysis of language is one of our most pressing needs. Machine translation has provided both a greater incentive and the means for accomplishing it.

An effective bar to thorough linguistic analysis lies in the complexity of language. The sound systems of most languages consist of no more than fifty entities. Determining these and their relations to one another is no child's play, but it can be done accurately and comprehensively with some effort. The forms of languages are much more numerous, but again are analyzable by the commonly used classificatory methods. But the syntactic and semantic units are virtually buried by their bulk. It is a considerable chore to analyze linguistically a text of several hundred thousand words, such as a novel, but if we did, we would still lack data on many of the six hundred thousand words listed in Webster's unabridged *Dictionary*. Even a complete analysis of Shakespeare's plays would give us data on fewer words than one in twenty-five. How are analysts to achieve more than an impressionistic description of English? Without repeating any of the painful and somewhat stultifying comparisons between the capacities of a man and a computer, we might simply agree that the only means lies in the exploitation of data-processing machines. To achieve a total description of English or any language would require their use, and the assistance of experts competent to manage them. The possibility of translation by machine has provided this opportunity. When adequate programs[2] are available, and when mechanical means of reading texts into machines are developed, we will be able to deal with huge quantities of linguistic material. This amount will increase progressively as machine translation is carried on, and may even begin to approximate the linguistic material any speaker has encountered in his lifetime; accordingly we may provide a scientific description as complete as our intuitive understanding of any given language.

Work in machine translation to the present has only modest achievements to report. Probably the most advanced procedure in question has been developed by the group working at Georgetown University under

[2] The word *program* has been adopted by computer specialists for the elaborate machine instructions devised to handle a problem or to manipulate data with a computer.

the direction of Professor Leon E. Dostert. In response to a request Professor Dostert arranged to have a new text in Russian supplied to a computer, and by means of the program which the Georgetown research group had developed the computer provided a translation which a professional translator at the Library of Congress considered adequate. Considerable publicity has been given to a procedure developed by the International Business Machines Corporation for the Air Force, to translate Russian into English. So-called translations produced by the procedure give no hope that it will achieve startling results, for it does little more than string along Russian elements in their English counterparts. A survey of the present status of work in machine translation would reveal that techniques now in operation are small in number, yet publication on theory and a variety of approaches is staggering. Any of the surveys entitled *Current Research and Development in Scientific Documentation* published by the National Science Foundation[3] illustrates the activity in computer research on language.

Much of the early, even much of the recent publication can be disregarded. For anyone who wishes to examine the studies produced there are several large bibliographies. One, which was planned by Russian scholars to be absolutely complete, uncovered such a tremendous amount of material that it has not yet appeared. Until it does, the *Bibliography of Mechanical Translation* by E. and K. Delavenay (The Hague, 1960) is quite adequate. For a brief summary of early work, and an introduction to the problems facing machine translation, one can scarcely do better than E. Delavenay's *La machine a traduire* (Paris, 1959). Yehoshua Bar-Hillel's "The Present Status of Automatic Translation of Languages,"[4] is another authoritative survey. Bar-Hillel's pessimism about the possibility that machine translation may be able to produce complete translations that require no post-editing may be illustrated by the sentence he used in his demonstration: "The box was in the pen." In working towards an approach which may yield translation of this sentence, we may use another, somewhat less troublesome sentence, and with it illustrate some of the problems facing machine translation.

If one were to translate *He was here yesterday* into German, the result might be: "Er war gestern da." To produce such a translation automatically, one would simply need to code each English word with an appropriate number, provide other numbers for the German words, and write a program to transpose the one set of numbers to the other.

[3] Number 7 was published in November, 1960.

[4] *Advances in Computers,* ed. by F. L. Alt (New York and London, 1960), pp. 92–163.

The program would take care of differences, such as putting German *da* after the equivalent for *yesterday,* by having a rule that all adverbs of time in German must precede adverbs of space. Obviously the necessity for this rule would require parsing the German and English words, not merely arranging a one-for-one substitution. The necessity for handling this problem of syntax and problems of morphology may indicate why machine translation requires, and fosters, complete linguistic analysis. Further illustrations could be supplied *ad infinitum;* they would only amplify the statement that we must make thorough linguistic analyses, and program into machines grammatical descriptions, not merely entities of one language with a likely translation to the other. The work involved may seem appalling; it also has a fascination similar to that exerted by any investigation on interested investigators, and it can be carried out. Put differently, a morphological and syntactic description of a language, though complex, can be accomplished with some exertion.

The semantic patterns of a language, however, require much more detailed descriptions. German *da* as adverb does not always mean "here." A statement like *Da stand er* we translate "There he stood." In other statements like *Da wusste ich nichts weiter* we translate "Then I didn't know anything further." We see that *da* may correspond to English *there, then, here;* in a good translation we must make the proper choice. Our approach to a solution may be that applied generally in linguistics; we define semantic entities not by their relation to the outside world but by their relation to other linguistic entities. We might establish a semantic rule, somewhat as we do syntactic rules, that whenever forms of *stehen* are used with *da, da* is to be translated *there,* not *here.* Such rules might require classifying many German verbs by their effect on the translation of *da.* This obviously would be a tremendous job, one not yet carried out explicitly. Yet every German-English bilingual of some competence has carried it out implicitly. It is precisely in the execution of such complex and painful compilations that a computer excels. To carry them out we merely need adequate computer programs; these can be produced after efforts similar to those expended on programs for dealing with problems in mathematics, in ordnance, in weather forecasting. The chief difference may lie in the novelty of carrying out minute and large-scale analyses of data from the humanities rather than the natural sciences.

Investigators interested in practical results from machine translation take some comfort in the relative simplicity of the language used in technical articles. In aiming at conveying their conclusions accurately, scientists use stereotyped expressions. The result may be ungraceful, but

the consequent consistency of usage brings far fewer problems than does everyday language, not to speak of highly varied literary languages. The frequent *Es sei* of German scientific texts, as in *Es sei die Strecke AB kongruent der Strecke A'B'* (Let the distance AB be congruent to the distance A'B') would be dealt with as a unit, and repeated as infallibly as it recurs in technical texts. A translation by machine of such stereotypes would be as accurate, and as dull, as the result produced by a professional translator, but much more rapid.

Devising means of handling texts which do not consist wholly of stereotypes provides some of the chief interest for workers in machine translation. Human beings use words with various meanings in varying situations, with few problems of misinterpretation. If the sentence *The compound was unknown to him* occurred in texts dealing with chemistry, we would have no trouble determining the meaning of "compound"; if it occurred in a treatise of life in China we would shift grounds quickly, and interpret it differently. A machine dealing with each sentence as an entity would be unable to make the shift. Programmers could of course prepare a program to yield two or more translations of "compound," to match its two meanings in the different contexts—and these would be automatically printed out whenever the word "compound" occurred; but the quantity of alternate translations necessitated by such a procedure would confuse the prospective client. Some procedure must be hit upon which will simulate man's activity when he confronts ambiguous sentences. It is in such simulation that work with computers will offer its greatest interest to the humanities. For to achieve adequate machine translation, ultimately computers will have to be programmed for the flexibility, the capacity for storage and the ability for learning which man possesses. In working to duplicate one of our most complex activities—transmittal of information through language—humanists for the first time have a means of dealing with their subject in accordance with techniques developed in the sciences: they will arrange experiments, form hypotheses, test them to determine whether their hypotheses were useful or correct, and if not, locate the cause of error. At last humanists will be able to concern themselves with the fundamentals of their discipline rather than with peripheral or trivial activities. Though they may claim for their present activities superiority to those of the scientists with their concern for inert or dissected substance, much of the activity of humanists is a type of engineering: production of texts, of bibliographies, of biographies. Other work is classificatory, such as comparing one poem, novel, painting with another, pigeonholing works of art. Often eminence in the humanities is achieved by mastery of a subtle and varied rhetoric rather than new

understanding of man. In computers humanists finally have a vehicle for dealing with their subject adequately, for constructing models of it, and accordingly for testing hypotheses at will. The point is not that like positivists with new potentialities they may record, classify, and rearrange myriads of facts, but rather that they will be able to build a model of a structure, a world. With this possibility they may simulate the "world of the poet," Wellek's crucial requirement for a "theory of literature."[5]

In keeping with the normal support for research, that providing funds for work in machine translation is aimed at a concrete goal: at managing the tremendous volume of material which is now published annually, by putting it in a different language, indexing it, making segments of it generally available. Although this aim seems remote from "poetics," from a "theory of literature," from dealing with "the world of objects which poetry and fiction build up in our imagination," it now appears that only through simulating such worlds can machine translation be successful. Kafka's world may be a "distortion of reality" but what are we to say about that of the nuclear physicist? Certainly it too is "highly selective" and to most of us fantastic. Yet if we are to achieve the greatly desired aim of accurately translating materials describing the world of the Russian, the German, the Chinese physicist into a different language, we will have to build a model of that world into a computer. If we achieve one such model, another will be simple; ultimately we will have models of Shakespeare's world, of Dante's, of Faulkner's, of Nietzsche's, even of Hitler's. As at present, these highly selective worlds will form the area of study of the humanist. Unlike the present he will not be confined to dealing with tiny segments or superficially with broad outlines, but he will have a means of mastering them in their entirety.

Consequences of this prospect need not be outlined in detail. If the world of Kafka can be created, so can that of a psychopath. When understood, the psychopath's world may be modified in somewhat the same way as that of a diabetic now is. Moreover, the world of any individual or group may be determined. We know little about the way of the chemist, the biologist, or any group which has subjected itself to a discipline and has as a result been modified. Construction of the theory of any science will be a much more useful contribution of the humanities to the sciences than is the attempted marriage in science history, which now is acclaimed the bridge between the two cultures. To achieve this aim the humanities will have to attract students other than those who

[5] *Style in Language*, ed. by T. Sebeok (New York, 1960), p. 411. Phrases in quotation marks in the following paragraph are taken from Wellek.

find science too difficult; these students must be provided with unparalleled training. For in handling the complex data of human behavior, mathematical manipulations will need to be developed that have only been foreshadowed, e.g., in probability theory. This new role of the humanities will follow the construction of adequate computers. Programmers are already at work on the requisite programs; the machines should be available within the decade. The humanities will then emerge from the tranquility of Gothic chambers to become again the central discipline.

*The University of Texas*
*Austin, Texas*