[From: Morris Halle et al.(eds.) *For Roman Jakobson* (The Hague: Mouton, 1956)]

SOME LINGUISTIC PROBLEMS IN MACHINE TRANSLATION*

BY PAUL L. GARVIN

THE GEORGETOWN-IBM EXPERIment in machine translation,[1] in which I participated, raised a number of problems of linguistic method and threw light on some important phases of general linguistic theory.

The purpose of machine translation (MT)[2] is to have a logical machine perform a task which so far has been performed by skilled human beings only - that of translation, that is, "the transference of meaning from one patterned set of symbols occurring in a given culture . . . into another set of patterned symbols occurring in another culture .. . ."[3]

Two questions have to be answered before MT can seriously be attempted: (1) what are the discrete steps involved in the process of translation; (2) how can these steps be stated in terms of the modus operandi of logical machines.

The steps in translation can be discovered by a detailed reconstruction of the translation process from a comparison of the original text with its translation, that is, by translation analysis.

The results of the translation analysis, in order to be compatible with the modus operandi of logical machines, must be stated explicitly and unequivocally, that is, with each logical step - including the obvious ones - spelled out in detail, and in terms of the yes-no decisions required by the binary operation of electronic circuits.

The above two aspects of MT research - the translation analysis, and the verbal statement of its results in acceptable terms – are the proper concern of the linguist, since they must be based upon

* I am indebted to A. C. Reynolds, Chief Engineer, Stromberg Time Corp., Thomastown, Conn., for discussing these problems with me, especially as regards the characteristics of logical machines and programming (see fn. 4). I am of course solely responsible for my argument and conclusions.

[1] See L. E. Dostert, "The Georgetown-IBM Experiment," in: William N. Locke and A. Donald Booth, eds., *Machine Translation of Languages* (The Technology Press, M. I. T., John Wiley & Sons, New York; Chapman & Hall, Ltd., London, 1955), pp. 124-35; and Jacob Ornstein, "Mechanical Translation, New Challenge to Communication," *Science* 122.745-8 (October 21, 1955).

[2] Standard MT terms are: "source language" and "target language" for the languages from which and into which the translation is made, "input" and "output" for information fed into and received from the machine.

[3] Dostert, *op. cit.,* p. 124.

a knowledge of the structures of the languages concerned. The third aspect, the translation of the linguist's statements into a detailed program[4] for a particular machine - be it a specially designed translation machine (which, to my knowledge, does not yet exist), or a general-purpose complex logical machine (such as IBM's 701 computer used in the Georgetown-IBM experiment) - is within the scope of the programming specialist with whom the linguist cooperates.

Translation analysis differs from linguistic analysis by both its subject matter and its objective. Instead of a corpus of forms, its data are a set of reconstructed operations. Instead of discovering the pattern underlying the corpus, its purpose is to make possible the duplication of the operations by a logical machine.

Let me now discuss the data of translation analysis - the translation process - in some detail.

The translation process consists essentially of two sets of operations: operations of selection, and operations of arrangement.[5]

The selection operations have to do with finding the suitable equivalent for each unit to be translated; I shall discuss later what these translation units are. The arrangement operations have to do with making sure that the translations of each unit appear in the output in such an order that the text as a whole is properly translated.

Selection and arrangement are trivial where one-to-one equivalence exists between source and target units, and where the order of units in the source language is identical with the desired order of units in the target language. In such cases, a simple look-up and matching procedure would be sufficient, and the machine would not be required to "make any decisions": the machine could store, in its memory device, units of the source language together with their single equivalents, and a text could be translated by matching each unit of the input against the corresponding source unit in the memory device, and furnishing the equivalent stored next to it in the output, in the same order in which the input was received.

The selection and arrangement operations include decisions when a source unit has more than one possible equivalent in the target language, and when the sequential order of units in the source language is not identical with the desired order of units in the target language. Then the machine, in order to translate adequately, has to make decisions as to which of the several equivalents to select, and decisions as to whether to retain a given order of units or alter it. This means that in addition to look-up and matching, a "decision program" must be included in the MT setup.[6]

Such a decision program has to meet two requirements: it has to enable the machine to recognize the "decision points," that is, the passages of the input text requiring translation decisions, and it has to furnish the "decision instructions" enabling the machine to execute the correct decisions.

To provide the information on which such a dual decision program can be based is the crucial objective of translation analysis. Translation analysis can achieve this objective to the extent to which it can establish the predictability of translation decisions in terms of the structural features of the source and target languages, and of the functional properties of language in general.

[4] "Programming" is the detailed keying of a logical machine for the performance of a required set of consecutive operations.

[5] Cf. Karl Bühler's two levels of organization ("Klassen von Setzungen"), choice of words ("Wortwahl") and sentence structure ("Satzbau"), *Sprachtheorie* (Gustav Fischer, Jena, 1934), p. 73, also Vilem Mathesius' "onomatology" and "syntax," "On some problems of the systematic analysis of Grammar," *TCLP* 6.98 (1936).

[6] It is, of course, possible to design a simple dictionary program in which the machine merely furnishes the one or several equivalents of each input unit from its memory in the order of the original text, to be worked into a viable translation by a knowledgeable human editor. Such a "mechanical dictionary" has been talked about in the MT discussion, but the Georgetown-IBM experiment was designed to test the feasibility of MT without either "pre-editing" or "post-editing" (Dostert, *op.cit.,* p. 134).

What, then, is the extent and nature of this predictability? Let me discuss the matter first in regard to selection decisions, and then in regard to arrangement decisions.

Selection decisions in translation concern meaning equivalence; to establish the conditions for the selection of a given equivalent, the various aspects of meaning involved here must be sorted out.

Every unit in the source language can be assumed to have a system-derived general lexical meaning[7] proper to itself. Let me define my terms: by system-derived meaning I mean that range of meaning which is proper to a linguistic unit by virtue of its place in a system of comparable units (in practice, the component of meaning recurrent in all of its distributions); by general meaning I refer to the total of this range of meaning, not merely to the most common and obvious segment of the range; by lexical meaning I designate the meaning uniquely proper to the given unit and not shared by any other unit in the system.

Every unit in the target unit must then likewise be assumed to have a system-derived general lexical meaning proper to itself. Since by definition two languages constitute two different systems, the range of meaning of no unit in one can be assumed to coincide exactly with that of a corresponding unit in the other.

This incomplete coincidence of system-derived ranges of meaning is the causative factor in the problem of selection: the range of meaning of a unit in the source language may include pieces of the range of meaning of several units in the target language, the result of which are the multiple equivalents referred to above. The scope of the problem varies for different areas of the lexicon: in colloquial or literary vocabulary there is usually much less coincidence than in technical vocabulary. Much of the latter is derived from the same international urban culture and therefore has reference to a similarly structured cultural reality, it furthermore has the relative exactness of reference required by the "intellectualization" of the standard language in technical and scientific discourse,[8] hence closer coincidence of the ranges of meaning can be expected. The selection problem can on this level be further reduced, though not completely eliminated, by more refined lexicography: most bilingual dictionaries, for instance, list more equivalents than would be necessary with a more careful matching of ranges of meaning (not to mention that equivalents are often poorly chosen).

The selection problem can then be rephrased as follows: what portion of the total system-derived meaning of the source unit applies in any given textual fragment, and to which of the possible equivalent portions of target units does it correspond?

The answer lies in considering the relationship between the linguistic unit and its environment. By virtue of its place in a system, every linguistic unit "brings with it" into each environment a system-derived range of meaning. This range of meaning is postulated to be relatively wide and relatively vague, since in practice it has to be abstracted from a multitude of environments, and has its value only by virtue of its opposition to other comparable units. The system-derived meaning is in every linguistic context and extralinguistic situation interpreted by the receiver in terms of this environment, and this reinterpretation - essentially a narrowing and specification of the system-derived meaning - can, in Karl Bühler's term (see fn. 7), be called field-derived meaning, in turn consisting of a contextually derived and a situationally derived component.

[7] The subsequent discussion is partly based on Karl Bühler's concepts of "feldfremd" *(*= system-derived) and "feldeigen" (= field-derived) *(Sprachtheorie,* p. 183), and on Roman Jakobson's treatment of general meaning and basic meaning, "Zur Struktur des russischen Verbums," *Charisteria Guilelmo Mathesio ... oblata* (Prague, 1932), pp. 74 ff., and "Beitrag zur allgemeinen Kasuslehre," *TCLP* 6.240 ff. (1936).

[8] For a discussion of this concept, see B. Havránek in *Spisovná čeština a jazyková kultura* (Prague, 1932), pp. 45-52, translated in "A Prague School Reader on Esthetics, Literary Structure, and Style," *Publications of the Washington Linguistic Club* 1.5-9 (1955).

What is found in a particular text to be translated is thus not the system-derived meaning as a whole, but that part of it which is included in the contextually and situationally derived meaning proper to the text in question. The equivalents of the total system-derived range of meaning can then be selected in terms of the applicable contextually and situationally derived portions of that range.

Selection decisions are thus, in terms of the above, contextually determined and situationally determined, and the objective of translation analysis becomes one of singling out the specific determining factors for each decision in the given context and situation.

This raises the extremely important problem of the boundary line between linguistic context and extralinguistic situation, since it is difficult if not impossible to envision a logical machine capable of extracting information from the multistructured extralinguistic environment of a text. It is obvious that all non-linguistic phenomena accompanying a textual unit are part of the extralinguistic situation; this does not, however, imply the converse, namely that all linguistic material accompanying the textual unit is part of the linguistic context in a technical, that is, operationally useful sense.

On the contrary, the distinction between linguistic and extralinguistic environment is functionally relevant only if it is made, not in terms of the substantive nature of the environmental material, but in terms of the relationship of this material to the textual unit under consideration. That is, the distinction must be made between environmental material linguistically related to the unit under consideration, and material not so related, whether it is substantively extralinguistic or not. This means that only a certain portion of the linguistic material accompanying a given textual unit can be considered its linguistic context in the technical sense; the remaining linguistic material, together with the extralinguistic material, technically is included in the extralinguistic situation.

Thus, the linguistic context can be defined specifically as all those textual units (i.e., linguistic material) that stand in a linguistic relation to the unit under consideration. A linguistic relation, to be meaningfully distinguishable from a non-linguistic relation in this connection, is defined as a necessary syntagmatic dependence.[9] The boundary line of the linguistic context then becomes the sentence, since the latter can be defined as a syntagmatically self-contained unit.[10]

Summing up the foregoing, the selection decision consists in singling out, within the total system-derived range of meaning of a unit, that portion of the range which applies to the environment in which the unit occurs, and in giving the translation equivalent corresponding to this range. That portion of the environment which qualifies as linguistic context is amenable to machine programming (since all necessary syntagmatic dependences can presumably be formulated in programmable terms), that portion which - whether of linguistic substance of not - has to be classed as the extralinguistic situation is not amenable to programming, since the relations linking this part of the environment to the textual unit in question include too many variables to be manipulable in the required precise terms.

Thus, the extent of machine translatability is limited by the amount of information contained within the same sentence, as defined above.

Arrangement decisions in translation concern the equivalence of sequential relations - the objective of an arrangement decision is to arrive at an order relationship in the output functionally equivalent to that of the input, as a result of which the original order is either retained or altered.

[9] Syntagmatic is here used in the sense proposed by André Martinet, namely in reference to relations within the same text. Dependence is used in the Hjelmslevian sense, as discussed in my “Delimitation of Syntactic Units,” *Language* 30.346-7 (1954).

[10] See my definition of the sentence in Ponapean as one of several mutually tolerant units, *loc. cit.,* p, 347.

To establish the conditions for this retention or alteration of order, the functions of the various sequential order relations in the source and target languages have to be established and compared.

We are thus again dealing with problems of meaning equivalence, but instead of the meanings of units we deal with the meanings of relationships, and instead of lexical meaning we are dealing with grammatical meaning, that is, meaning not unique to each given unit but shared by more than one unit. While selection decisions are thus unique and specific to each unit, arrangement decisions are recurrent for classes of units.

Since, however, ranges of grammatical meaning can no more be expected to coincide from one language to another than can ranges of lexical meaning, every arrangement decision contains within its scope a possible selection sub-decision, determining which of several alternative sequential order relations is to be executed.

What has been said about the contextually derived and situationally derived determination of selection decisions holds equally, of course, for the selection sub-decision within the arrangement decision, although we may for all practical purposes assume that, since grammatical meanings are linked to grammatical patterns, contextual determination will here far outweigh situational determination. We may furthermore extend the scope of this contextual determination to cover the entire reach of the arrangement decision, not merely its decision aspect.

The above means that, unlike selection decisions, arrangement decisions can - at least as far as the purely linguistic factors are concerned - be considered entirely within the bounds of machine translatability, since the limits of the linguistic context need not be exceeded.

So far, I have discussed the matter only in terms of the operations involved in the translation process, without analyzing the translation units that are manipulated in these operations.

Let me now turn to translation units.

This part of the MT problem area concerns not only the translation units themselves, but also the very crucial problem of the relationship of translation units to sensing units, that is, the kind of units which a logical machine is capable of sensing.[11]

I shall therefore discuss the properties of translation units and attempt to relate them to the corresponding characteristics of sensing units.

First of all, translation units are linguistic units, that is, they are recurrent partials which can be separated out from a text by procedures of linguistic segmentation. Linguistic segmentation - as has lately been discussed in the literature[12] - is based on criteria of both form and meaning; sensing, on the other hand, occurs in terms of form (i.e., physical impulses) alone. This means that where homonyms or homographs, for instance, constitute separate linguistic units (and hence separate translation units), all formally identical homonyms constitute one and the same sensing units, resulting in an additional selection problem.

Secondly, although translation units qua linguistic units are discrete, their physical manifestation need not be discrete. That is, while printed matter, for instance, is composed of discrete letters, speech or handwriting is at least in part continuous, with the physical continuum being interrupted only at some of the linguistically present unit boundaries. In order to sense speech or handwriting, a machine would have to be equipped with an elaborate resolving device capable of breaking up the continuous portions of the physical phenomenon into discrete units. Even then, at the present state of engineering, such a device can only accommodate constant continua, but not variable continua such as normal speech or handwriting.[13] This means that the input of MT, to be

[11] Sensing is the handling of the input data by the machine.

[12] For the most recent discussion of this problem, see Henri Frei, "Critères de delimitation," *Word* 10.136-45 (1954).

[13] This assertion is based on a statement of current engineering opinion.

practicable in the foreseeable future, will have to be limited to print or some other physically discrete source.

Third is the key problem of the respective extension of sensing units and translation units.

Let me elaborate. Sensing occurs in simple linear progression. For purposes of MT, a machine can be programmed to sense one letter at a time, and to recognize spaces as word boundaries - individual printed words separated by spaces will thus become the sensing units[14] which can be matched against comparable units stored in the memory.

In terms of the translation operations applied to the input, however, one printed word separated by spaces does not necessarily constitute a single translation unit. The translation unit - be it a selection unit or an arrangement unit - may consist of one word, or of less than one word, or of more than one word, depending on how much of the text is subject to a given operation at a given time. Thus, there is no one-to-one correspondence between sensing units and translation units, and additional programming has to be introduced to convert sensing units into translation units.

To determine the specific size of translation units, a text must be subjected to linguistic segmentation in much the same way as in determining the extension of any linguistic unit: defining criteria for each type of unit have to be formulated, such that the application of these criteria will result in an exhaustive segmentation of the text into units of that type.

Translation units, as was intimated above, are of two types, in terms of the two types of translation operations: selection units and arrangement units. In operational terms, they can be defined as follows: a selection unit is a unit, the total translation of which is not the sum of the translations of its parts; an arrangement unit is one which, as a whole, occupies a certain sequential order position and the translation of which, as a whole, is subject to a possible shift in sequential order position. In linguistic terms, selection units correspond to lexical units, arrangement units correspond to morphemic units, defined as follows: lexical units are units, the meaning of which is not predictable from the meanings of component units (if such can be segmented out); morphemic units are units, the distribution of which can not be stated in terms of component units (if such can be segmented out). Both lexical and morphemic units may consist of one or more linear segments such as morphemes or words; their internal structure is irrelevant from the standpoint of their meaning or distribution as wholes (or, in MT terms, from the standpoint of the selection and arrangement of their equivalents in translation).

In summary, the properties of translation units (shared in a broader sense by linguistic units in general) contrast with those of sensing units as follows:

(1) translation units are defined by both form and meaning, sensing units are defined by form only;

(2) translation units can be manifested in a variable continuous physical substance,[15] sensing units only in a discrete, or perhaps a constant continuous, substance;

(3) translation units can not be segmented in simple linear progression, sensing units can be sensed only in simple linear progression.

I have tried in the above to formulate some of the considerations which, as a linguist, I regard as basic to MT. They are founded on some assumptions about the functioning and properties of

[14] I am disregarding punctuation marks as possible sensing unit boundary markers, since the problem of varying extension is of the same order for such larger units as might be delimited by punctuation marks, and the subsequent discussion, mutatis mutandis, therefore applies to them as well.

[15] Louis Hjelmslev, *Prolegomena to a Theory of Language,* Francis J. Whitfield, transl. (= IUPAL Memoir No. 7, Baltimore, 1953), p. 50.

linguistic systems which are inferential in nature, since after all only the noises of speech or their graphic representation (la parole) are directly observable, but not the underlying structure (la langue).

Machine translation involves, in essence, the designing of a machine analog to the postulated resultant of the interaction of two linguistic systems - it will thus allow an instrumental verification of the theoretical assumptions about the nature of these systems and the nature of their interaction, such as the postulated relational difference between the linguistic and the extralinguistic, and the postulated non-linearity of linguistic units.

As a by-product, the requirement of explicitly and unequivocally formulated statements acceptable for programming purposes, will contribute to the rigor and scientific logic of linguistic analysis.

INSTITUTE OF LANGUAGES AND LINGUISTICS
GEORGETOWN UNIVERSITY