

[From: Jacob L.Mey (ed.) *Language and discourse*  
(Amsterdam: John Benjamins, 1986)]

PAUL L. GARVIN  
Buffalo, New York

#### SEMIOTIC ASPECTS OF MACHINE TRANSLATION

There is a trivial way of looking at this question: machine translation belongs into semiotics because semiotics deals with sign systems, language is a sign system, and machine translation deals with language. Once this has been said, what can be said next? In other words, are there nontrivial connections between machine translation and semiotics?

In my view, there are at least two such connections; they are connections between the theoretical foundations of machine translation and aspects of semiotic theory. One of these has to do with general sign theory and how it relates to and can be applied in machine translation; the second concerns the possibility of relating the problems of esthetic translation by machine (albeit only experimentally) to some theoretical principles of esthetic semiotics.

#### 1. GENERAL SIGN THEORY

From the beginning of my interest in machine translation (see Garvin 1956: 182-4) I have always felt that one important area of semiotic theory had direct bearing on machine translation, namely, Karl Bühler's doctrine of the two fields (*Zweifelderlehre*, 1934: 119). Bühler

divides the general environment in which signs function, their surrounding field (Umfeld, 1934: 52), into a symbolic field (Symbolfeld, 1934: 149-154) and a deictic field (Zeigfeld, 1934: 79-82). The symbolic field consists of those other signs of the same system to which the sign under consideration is directly related - in the case of language, the strictly linguistic context. The deictic field includes all the remaining context - in the case of language, what might be considered the extralinguistic setting. The question which then arises and which is the most important from the standpoint of machine translation concerns the boundary between the two fields. The importance of this boundary derives from the fact that the relations which a sign has to other signs of the same system with which it is connected are of a different kind than the relations that a sign has to its wider environment - that is, relations within the symbolic field are of a different kind from those within the deictic field. The significance of this difference for machine translation is that this difference in the relations requires a different kind of processing for the two kinds of fields.

The question of the boundary between the two fields can for machine translation purposes be resolved by assuming that the symbolic field is limited to the particular sentence (in the broad sense, including compound sentences) in which a given linguistic unit is contained. This is based on the underlying further assumption that properly linguistic relations in the strict sense are limited to the sentence: relations between sentences (in spite of such well-known phenomena as anaphora) are not of the same strictly linguistic

sort (e.g., the antecedent of an anaphoric pronoun may be a picture). From a machine translation standpoint it can further be said that relations within the sentence are deterministic in nature, i.e., capable of being processed by more or less an algorithmic process based on parsing. Relations beyond sentence boundaries, on the other hand, tend to be probabilistic. This means that algorithmic processing will not suffice and other kinds of processing (such as possibly heuristic) will have to be considered; much less is known about the latter than the former. This means that the resolution of ambiguities based on conditions that transcend sentence boundaries will be based on different and less clearcut principles than that for which conditions can be ascertained within the sentence. The latter can to a large extent be resolved on the basis of an adequate parse of the sentence, while the former will require more complex and less clearcut probabilistic solutions.

This can be illustrated by the lexical translation ambiguity presented by the Russian word работа which is rendered in English as either "work" (work accomplished, as on the job) or "paper" (a paper written, as in a scholarly journal).

In example (1), the linguistic context (following

(1) Работа В. И. Иванова на эту тему ...  
paper/work (by) V.I. Ivanov on this topic ...

genetival modifier consisting of a nominal structure identifiable as a proper name by virtue of its graphic shape) serves to resolve the translation ambiguity and to select "paper" as the appropriate translation. In

example (2), on the other hand, the translation

(2) В. И. Иванов написал статью на эту тему.

V. I. Ivanov wrote (an) article on this topic.

Эта работа ...

This paper/work...

ambiguity is not resolvable by just parsing either of the two sentences, but only on the basis of drawing upon the relation between the two sentences: the word статью in the first of the two sentences of which the example consists constitutes an antecedent of sorts of the word работа contained in the second sentence. This relation is not identifiable on the basis of parses but might be dealt with by some probabilistic mechanism such as, perhaps, an "antecedent storage" - a running store of the semantic categorization labels of all the relevant terms (however 'relevant' is defined for a given translation system) encountered in the sentences already processed, on the basis of which a resolution of the translation ambiguity encountered in the current sentence can be attempted. Thus, in the case of example (2), terms such as the already noted статью, as well as написал, encountered in the immediately preceding sentence (and presumably other terms in earlier sentences) will carry semantic categorization labels favoring the choice of a translation such as "paper", thereby resolving the ambiguity.

A similar ambiguity resolution, but with choice of the other translation alternative, can be suggested for examples (3) and (4).

- (3) Работа в этой фабрике ...  
paper/work in this factory...
- (4) При стройке метро несколько  
during construction (of the) metro several  
раз был перевыполнен план. Работа...  
times was overfulfilled (the) plan. Paper/work...

In example (3), the linguistic context (following prepositional modifier consisting of a prepositional structure attachable to the governing noun through an appropriate government code) serves, as did the linguistic context in example (1), to resolve the ambiguity and to select the appropriate translation - this time, "work" and not "paper". In example (4), as in example (2), the ambiguity again is not resolvable on the basis of parsing either of the two sentences but only by drawing upon the relation between them; the words *стройке* and *план* in the first of the two sentences constitute antecedents of sorts of the word *работа* with which the second sentence begins. Once again, this relation will require something like an "antecedent storage" for its identification, in which terms such as *стройке* or *план* will carry the semantic categorization labels allowing the selection of the translation alternative "work" rather than "paper", thus serving to resolve the ambiguity.

The antecedent storage could be combined with some wider-reaching technique such as information-retrieval methods serving to establish the general semantic character of a text (e.g., some form of key-word-based

processing), especially for those cases of ambiguity where the immediately adjacent sentences do not provide the needed information.

The link between this problem area and some of the current arguments about the semantic aspect of linguistic competence versus "real world knowledge" is evident.

## 2. ESTHETIC SEMIOTICS

The consideration of machine translation of esthetic text (such as literary passages) will to many appear as a contradiction in terms: how can a machine program be expected to perform (or simulate) an essentially creative activity? Thus, in order to consider the possibility of machine translation - or, more realistically - machine experimentation - in this area, it will therefore be necessary first of all to find an approach to the study of esthetic text which does not posit creativity (or beauty, or some other noncomputable property) as the primary defining criterion of esthetic phenomena.

Such an approach could well be based on Jan Mukařovský's concept of foregrounding as the basic explanatory principle of esthetics. This concept is based on the notions of the expected as over the unexpected - a conception which does not require recourse to the non-computable properties noted above. Foregrounding is the unexpected (unexpected behavior, unexpected features of objects), highlighted against a background of automatizations which are the expected (expected behavior, expected features of objects; cf. Mukařovský 1932: 19). Foregrounding may occur under ordinary, everyday circumstances in an unstructured manner (as in spontaneous joking and punning); in esthetic texts and other

esthetic objects, *it* occurs in a highly structured and systematic manner (such as the highlighting of certain linguistic features, for instance rhyming phonological elements, in a text). In Mukařovský's terms this constitutes the difference between the structured and the unstructured esthetic (see 1948: 31-32).

In the processing of text, the difference between the expected and the unexpected seems to me to be computable, at least in principle, since under the expected one can include the kinds of features, terms, expressions and constructions most commonly found under certain definable conditions. One can then assume that anything other than the latter would be unexpected and hence foregrounded.

This is particularly applicable to the structured esthetic, since the unexpected phenomena, the foregroundings, will here occur systematically. Recognition routines can then be envisioned that will identify such structured foregroundings, based on both formal linguistic properties, and (though undoubtedly less easily) on semantic ones.

An example of the machine recognition of formal properties might be the identification of alliteration or rhyme. Such poetic devices could conceivably lend themselves to identification through some phon signed simply to "spot" foregrounded passages, that is, a recognition routine such as that just discussed, without a command routine for producing translated equivalents. One way of designing such a "foregrounding spotter" would be to design a routine for identifying expected structures and the assuming that foregrounding is present whenever the routine fails because of the unexpectedness

of the structures encountered. This could be combined with an interface capability allowing a human investigator to examine the text and the results of its machine processing, and to give his interpretation.

Another point made by Mukařovský is relevant here, namely, that foregrounding occurs against two automatized backgrounds. In the case of esthetic text, one of these backgrounds is ordinary language usage, the other is the esthetic canon of the period (1932: 22). I am inclined to make his statement more specific and characterize the second background as the esthetic canon of a given readership or audience. The machine analysis of both kinds of backgrounds is conceivable (although, need I say, not immediately implementable): a known esthetic canon can, in principle, be incorporated in a recognition routine just as can a known linguistic usage.

Thus, the role of machine translation research here will clearly not be one of contributing directly to the achievement of literary translation, but - as I have already noted - one of serving as a testing ground for crucial esthetic semiotic notions.

#### REFERENCES

- Bühler, K. 1934. *Sprachtheorie*. Jena: Gustav Fischer.  
 Garvin, P. L. 1956. "Some linguistic problems in Machine Translation. For *Roman Jakobson* ed. by H. G. Lund et al. The Hague: Mouton.  
 Garvin, P. L., ed. & transl. 1964. *A Prague School Reader on Esthetics, Literary Structure and Style*. Washington, D.C.: Georgetown University Press.  
 Mukařovský, J. 1932. "Jazyk spisovný a jazyk básnický". *Spisovná čeština a jazyková kultura* ed. by B. Havránek & M. Weingart. 123-149. Prague; quoted from Engl. transl. *Standard Language and Poetic Language*,



- in Garvin 1964. 17-30.
- Mukařovský, J. 1948. "Estetika jazyka". *Kapitoly z české poetiky*, 2nd ed., vol. I, 41-77. Prague. Quoted from Engl. transl. *The Esthetics of Language*, in Garvin 1964, 31-69.
- Mukařovský, J. 1972. "Vančurovská prolegomena". *Cestami poetiky a estetiky*. 242-245. Prague.