American Journal of Computational Linguistics Microfiche 9: 28

O P I N I O N

A RESTRICTED SUBLANGUAGE APPROACH TO HIGH QUALITY TRANSLATION

Victor Raskin
Institute of Philosophy
Hebrew University
Israel

This paper deals with an approach to the problems of automatic high quality translation and, more generally, of automatic language data processing, based on the restriction of the input of MT and other systems to a certain type of sublanguage. The approach was proposed by the present author in the framework of a general theory of sublanguages (see Raskin, 1971) and subsequently tested and used by his own and other groups of researchers in the USSR on the material of diverse restricted sublanguages (see ibid; Gorodeckij and Raskin, 1971, Pt.1). The paper consists of two parts. Part 1 contains a very brief exposition of the basic principles of the approach. In Part 2 some advantages of the approach over other (unspecified) approaches unrestricted in this way are mentioned in the context of a few important problems of high quality translation. Since these problems were also discussed by the contributors to *Feasibility Study on Fully Automatic High Quality Translation*, at certain points of Part 2 the paper enters a dialogue of a sort with some of them (all the quotations and references followed by a name only are taken from the contribution by the corresponding author to the said report).

1. SOME BASIC PRINCIPLES OF THE RESTRICTED SUBLANGUAGE APPROACH

It might be observed that in most cases when the practical need of constructing an MT system arises, its input, i.e. the linguistic material which is to be subjected to such treatment, is highly restricted by certain conditions: it is usually a relatively *narrow field of science or technology* with texts which are *relatively homogeneous*, with a *limited vocabulary*, a *restricted set of syntactic constructions*, a highly *structured substance of the content plane*, and a relatively *constant system of values for all the relevant pragmatic parameters* which are determined in this case not by the individual properties of any particular situation of communication, as is usually the case in casual communication, but rather by the position of the field itself among the contiguous fields as well as in non-linguistic reality, in general. For such restricted sublanguages a simple algorithm of automatic processing was constructed and proved to be highly efficient in its practical applications.

The algorithm is based on an over-important property which follows, logically and practically, from the features which characterize the class of restricted sublanguages in the theory of sublanguages, including those which were emphasized above and which result in the irrelevance of all surface structure distinctions among sentences with identical deep structure or the exact synonymity of all paraphrases (and, in fact, even near-paraphrases). This property implies that each stem in the vocabulary of a restricted sublanguage tends to play certain permanent role in

all the situations described by those sentences where the stem occurs, no matter whether it takes the form of a verb, noun, or any other part of speech. A minimal sufficient inventory of these roles, which are given the status of semantic characteristics of stems, is compiled (usually it does not exceed 15 items) and each dictionary entry is assigned a certain characteristic. Then a scheme of the maximally extended sentence of the restricted sublanguage, a maximal deep structure of a sort, is postulated in such a way that each sentence (or rather, each clause) can be represented as a (partial) realization of this structure. Such structures can embed, nest in, etc., each other. The dictionary of the restricted sublanguage with all its entries being assigned semantic characteristics and the schema of the maximally extended sentence of the restricted sublanguage are the two instruments on which the universal-algorithm is founded. Texts of the restricted sublanguage constitute its input, the output being a sequence of (partially) filled, ordered and subordinated schemes/deep structures. By making the semantic characteristic assigned to each stem of the restricted sublanguage, more or less detailed, one may control the depth of semantic analysis. With its subalgorithms of "ellipsis analysis", "homogeneous parts analysis", "boundary analysis", the algorithm operates as a universal Turing machine in the sense, that having been fed the universally standardized information on a particular restricted sublanguage, it proceeds to analyze it in the universal way and is equally applicable to each and every restricted language.

Is Restricted Sublanguage Approach, RSA, applicable to all, or at least most relevant cases or can it be applied only in a few exceptional situations? It has been argued elsewhere (see Raskin, 1971; Ch.4.1) that the first alternative holds true while in the cases in which a polythematic informational system is needed it seems worthwhile to treat the processed texts as belonging to several distinct restricted sublanguages, and after distinguishing them with the help of a not too complicated device, to make use of the technique developed for restricted sublanguages.

2. RESTRICTED SUBLANGUAGE SOLUTIONS TO SOME PROBLEMS OF HIGH QUALITY TRANSLATION

Semantics and pragmatics and the quality of translation.

Recent developments in semantic and syntactic theory have demonstrated the practically indefinite potential depth of a complete linguistic description which seems to require much scarcely accessible (at present, if not in principle) information on "speech act conditions, conversation rules, and semantic interpretation which must be associated in an idiosyncratic way with the lexical item in question", on "a theory of illocutionary acts", on "a theory of discourse which relates the use of sentences in social and conversational situations", and on "a theory of natural logic" (Fillmore), while the pragmatic dimension of the text is said to include answers to such heterogeneous questions as "by whom the text was produced, for which kind of audience it was meant, which kind of background knowledge the producer of the text assumed to be

available to the audience, the time, the place, and other parameters of the situation in which the text was produced, etc." (Bar-Hillel).

Now, it is obvious that for an adequate translation, no matter whether it is human or automatic, all this highly complex information must be obtained and taken into consideration, otherwise the quality of translation falls down sharply. It is equally obvious that all this is far beyond the linguist's reach at the present stage of linguistic knowledge.

In order to arrive at a practical solution of this problem one should impose some restrictions on the process of MT. In other words, certain criteria of the quality of translation should be formulated, and if necessary and possible, lowered. One might try to restrict the output of an MT system in the sense that it should certainly not produce what the user does not actually need It is evident that the user of a translation of a scientific or technical text will not require as much finesse and subtlety as the user of a translation of a literary text. Some (e.g. Garvin) are prepared to go even further and construct systems which would produce clearly inadequate though still tolerable translations (in a sense nobody has even succeeded in defining) in order to gain in speed. Now, when "machine-aided translation" or similar approaches are suggested, a restriction is imposed on what the computer is supposed (and thought of as capable) to do.

The restriction on the input in RSA determines, of course, some restrictions on the output (but, certainly, not to the extent

of tolerating barely acceptable translations). On the other hand, rather on the contrary, the simplicity and easier observability of the material of a restricted sublanguage make automatic translation feasible, allowing at the same time and for the same reason for the total accountability of the sublanguage which makes it possible to account for and use for the practical purposes of translation all the complex semantic and pragmatic information which might be relevant for translation. Of course, what makes it possible is that the degree of complexity of such information in the restructed sublanguage is very much inferior to what might be observed in language as a whole. What follows, however, is that restricting the input of an MT system to a sublanguage of a certain type RSA ensures high quality translation within the sublanguage and no further restrictions or lowering of the quality of translation is necessary.

It should be mentioned at this point that RSA shares with "machine-aided translation" the property of requiring a limited amount of predetermined and routine human participation prior to automatic processing.

Syntax and semantics, lexicon and grammar. One of the major claims of RSA is that, at least in applications to restricted sublanguages, intricate and labor-consuming syntactic algorithms (cf. Mel cuk, 1964) are redundant. The universal algorithm is based on semantics and is designed to use linguistic information of "lower" linguistic levels (viz. morphology and syntax) in a few exceptional

cases of semantic ambiguity. This emphasis on semantics rather than on syntax in automatic language data processing systems takes on a new value when compared to current discussions of the relations of syntax and semantics in linguistic theory and the existence of a clear-cut boundary between them. Probably influenced by the tendency, at present prevailing statistically in theoretical linguistics, to claim the priority of semantics over syntax, and, moreover, to negate the existence of the boundary, even those researchers in MT who do not seem to be influenced by RSA also speak in favor of such a "semantically-based" position (e.g. Mey). The latter position is indirectly reinforced by the fact that purely syntactic contributions to the *Study* (e.g. Petrick) fail to prove their pertinence to the problem of actual realization of MT bearing instead on the relation of recent theoretical innovations to the feasibility of MT (see below).

Thanks to its basic principles and internal organization RSA came independently to a justification of the claim made by Garvin that it is operationally more effective to delegate most of the grammatical information used in an MT system to the lexicon rather than to the parsing algorithm.

Linguistic theory and feasibility of MT. RSA seems to contribute to the solution of the major dilemma concealed in this phrase by providing, in a way similar to the one discussed above in connection with semantic and pragmatic problems, an interesting halfway position, a middle ground of a sort which in a sense combines

some relevant properties of the two extremes.

In the light of quite a number of promising developments and achievements in linguistic theory, the pertinent question is whether these have, do, or will, contribute anything to MT, or the latter, as Lyons thinks, "will neither contribute very directly to, nor depend very directly upon, advances in linguistic theory." This basically defeatist position has at least two aspects, the one being that language is claimed to be too complicated to be successfully subjected to automatic processing, an opinion many theorists would subscribe to, and the other, proclaimed by MT operationalists" (e.g. Garvin) that much of what has been recently proposed in grammatical and semantical theory is far too strong for MT, and much weaker models, as a possible theoretical basis for practically feasible MT are required. The latter consideration is interestingly illustrated by the fact of the recent emergence of working automatic systems of language data processing, quite close in their restrictiveness to RSA though, rather contrary to it, not aiming at theoretical generalization, which use "analysis-based grammars" (cf. Winograd, 1972).

-However, it is natural for the linguist to be suspicious of any attempts to base an MT system on a theory or a model, which has been demonstrated to be inferior to some other theory or model. Any serious attempt to make use of any linguistic knowledge for any purpose must, he might feel, be based on an adequate theoretical framework, otherwise the ever present danger of ad hoc decisions could hardly be avoided. What might be missed in this

argumentation is the fact that, when dealing with computerized applications of linguistics, we impose on the linguistic material a fundamentally different phenomenon, with laws and logic of its own, which may be very foreign to the nature of human language and the mental mechanism which underlies it, and this might force us to give up purely linguistic theories, even if they seem based on the properties inherent in man's nature, and to adopt, in man-machine partnership, a compromising approach which would account for both human and machine nature. It is not unimaginable, though rather distressing if true, that, due to the essential difference between the two, no linguistic theory claiming or exhibiting the property of adequacy to the nature of human language can be directly "computed", i.e. taken in by the computer;

It seems, and this is substantiated by the material of some papers contributed to the *Study* (e.g. Karttunen), that the more dependent on some recent development in "pure linguistics" a paper is, the less pertinent to MT it becomes. The contradiction between linguistic theory aiming at adequacy and practical needs of MT and, for that matter, other problems of computational linguistics, is self-evident. In this situation RSA seems to be doing a unique job of reconciling the two extremes, since on the material of a restricted sublanguage it might turn out that the application of a grammar based on adequate linguistic theory would be quite practical and there would not be any need to seek more feasible ad hoc solutions. Besides that, RSA may contribute a great deal to what is essentially

an issue between "theory" and "practice" by:

(1) providing a suitable "testing ground" for various conflicting theories or models, both for those which claim linguistic adequacy and analysis-based ones;

(2) allowing one to select the most preferable alternative on the basis of complete and easily accessible evidence which might be relevant for the choice;

(3) enabling one to limit the strength of a too powerful but valid theory or model by making suitable modifications on the basis of easily observable linguistic material of the restricted sublanguage.

The basic principles of RSA make one think of its language independence.

REFERENCES

Gorodeckij, B. Ju., and V. V. Raskin, 1971. *Methods of Semantic Investigation of a Restricted Sublanguage*, Moscow: Moscow University Press (in Russian)

Mel'cuk, I. A., 1964. *Automatic Syntactic Analysis*, Novosibirsk USSR Academy of Sciences Press (in Pussian)

Raskin, V. V., 1971. *Towards a Theory of Linguistic Subsystems*, Moscow: Moscow University Press (in Russian - an English translation is in preparation by Mouton)

Lehmann, W. P., and R. Stachowitz, 1971. *Feasibility Study on Fully Automatic High Quality Translation*, Griffiss Air Force Base, Rome Air Development Center, RADC-TR-71-295

Winograd, T., 1972. *Understanding Natural Language*, New York: Academic Press