GRAMMATIC AND SEMANTIC NORMATIVITY OF LINGUISTIC UNITS AND FEATURES AS A FACTOR OF AUTOMATIC TEXT PROCESSING

Z. M. Shalyapina

Institut vostokovedenija AN SSSR,
ul. Ždanova 12, 103 777 Moskva GSP, SSSR

All systems of automatic text processing are explicitly or implicitly based on two general linguistic assumptions: the assumption of grammaticality of the texts processed, and the assumption of their meaningfulness. These assumptions, however, cannot be considered as absolute laws: it is not uncommon that a text, though acceptable to most speakers of the corresponding language, still contains some morphologic and/or syntactic ingrammaticalities or cannot be completely interpreted in terms of "standard" semantics; and conversely, starting from an acceptable (meaningful) semantic structure one may as often as not fail to find fully "grammatical" language means that could express this structure with absolute accuracy (one of the usual translation difficulties). This is due, from our point of view, not only to incompleteness of linguistic and extralinguistic knowledge of separate people or to imperfections in the corresponding formal models, but also to the following two fundamental principles of linguistic performance:

1) a large number of requirements on lexico-grammatic (superficial) manifestation of natural-language texts, and on their semantic interpretation, are relative in that they characterize certain manifestations or interpretations as more or less normative (preferable) in the given conditions, rather than obligatory vs. inadmissible in the absolute sense;

2) the interaction between the requirements of grammatical and semantic normativity of texts adheres to a sort of complementarity principle: if the basic meaning of a text fragment is supposed by its author to be sufficiently transparent or known apriori to the text addressee, the grammaticality requirement for this fragment's surface manifestation may be somewhat slackened; if on the contrary, the author believes the text to contain much important information new to the addressee, the language rules used in composing its surface manifestation are apt to be as standard and rigid as possible.

In this presentation we intend to describe one way of incorporating the above principles in the design of the analysis and synthesis (generation) components of an automatic translation system.

The general structure of the system viewed from this standpoint is planned to be as follows.

The major part of factual linguistic information is formulated in the system regardless of the tasks of analysis or synthesis. It is shaped principally as a set of descriptive rules arranged into dictionary and grammar according to the so-called lexicographic principle and classified into two main types: the context-representation rules making up the contextual dictionary and grammar component, and the context-contrastive rules forming the inter-contextual grammar.

The rules of both types describe the possible superficial manifestations and semantic interpretations for elementary potential components of text structure. The kind of text structure serving as the point of reference for this description is defined in our model at the language-sign (LS) level, based primarily on the Saussurian conception of linguistic sign and roughly corresponding to the level of N.Chomsky's deep structures.

The context-representation rules proceeding from this type of structure specify the contextual functioning of

language units and features isolated at the LS-level, by relating them to their associated manifestations and interpretations. Essentially, they amount to statements of the following pattern: "If the LS-structure of a text contains a certain unit or feature X in a certain contextual position, this unit or feature can be superficially manifested (resp. semantically interpreted) in this text through the use of expression means Y (resp. of meaning constituent Z)."

The above principle of "relativity" is incorporated in these rules by supplementing each of them by its priority coefficient showing the degree of its normativity. In contrast to many other "preferential" linguistic models we emphasize the linguistic significance of these coefficients which, in our view, must be derived primarily from the interplay of synonymy and homonymy as phenomena inherent in natural language. With our linguistic description centered as it is around the notion of linguistic sign in the Saussurian sense, it is possible to evaluate these phenomena, as well as the priority coefficients required, in terms of statistical data bearing on the occurrence rate (relative frequency) of various specific manifestations and interpretations of each LS-structure among their alternatives.

The context-contrastive rules implement the "relativity" principle even more immediately. Their general pattern is: "If a fragment of the LS-structure of a text has several alternative manifestations (resp. interpretations) differing in a certain characteristics Y, preference should be given, all things being equal, to the alternative where the value of Y is related to the values of the same variable for the other alternatives in a definite way". In terms of such rules one can state all those particulars of the surface and/or semantic arrangement of natural-language texts (or of a special type of texts) which involve a kind of overall stylistic comparison, rather than the properties of individual linguistic units and structural features.

The descriptive part of the system is made operational by means of special control components acting as "planners" of the analysis and synthesis processes. One of the main tasks of these components within the framework outlined consists in grading the alternatives obtained from processing separate text fragments, as more or less promising for accomplishing the analysis (resp. synthesis) of the whole text, this gradation based, among other things, on the priority coefficients of the rules used to form (or check) different alternatives, and on the interrelation between these rules with respect to the grammar and semantics "complementarity" principle. In as much as this aspect of processing is concerned, the approach accepted makes it possible to re-interpret the well-known idea of "analysis through synthesis" and "synthesis through analysis" from the "normativity" angle of view. Thus, for analysis one can reduce this idea to a formalization of the following line of reasoning (quite popular with translators or people somehow concerned with texts in foreign languages): "Expression X in the text at hand cannot mean Y because had the author meant Y he would have much rather used expression Z".

Apart from affording better processing accuracy and efficiency, explicit introduction of data on normativity and preferability of linguistic units and features throughout all the major components of a text processing system, and drawing on statistical characteristics of LS-units' contextual manifestations and interpretations as the controlling factor in selecting the more "promising" among the alternative routes of processing concrete texts, seems to have one more asset. Namely, possibilities are opened up for automatically perfecting the system's functioning when required, and tailoring it to different text styles, by way of modifying the priority coefficients of the linguistic rules involved, directly from the current results of the system's operation.