PROBLEMS OF FORMAL REPRESENTATION OF TEXT STRUCTURE FROM THE POINT OF VIEW OF AUTOMATIC TRANSLATION

Z.M. Shalyapina

Institute of Oriental Studies
of the Academy of Sciences of the USSR

Moscow, USSR

Summary

The paper is devoted to linguistic problems of defining the basic formalized representation of text in an automatic translation system within the framework of the so-called integral formal model of the translation process, the primary requirement for this representation considered to be a compromise between its semanticity, superficiality, and exhaustiveness. A representation covering five major aspects of text structure (its lexico-grammatical composition; its predicate-argument organization on the semantico-syntactic level; the syntactic grouping of its units; the anaphoric relations between them; the peculiarities of their linear arrangement) and referred to as Combined Structural Representation (CSR) of text, is described to show the ways and means of achieving this compromise in the Japanese-Russian Automatic Translation Project, now under development at the Institute of Oriental Studies of the Academy of Sciences of the USSR (Moscow).

Introduction

Many problems of the automatic processing of text require for their effective solution a previous analysis of the text processed, aimed at transforming this text into its intermediate formalized representation of some kind, more suitable for further processing than the text itself. When determining the concrete characteristics of such a representation one must obviously take into account the operations meant to be applied to it, or to be performed on its basis within the framework of the system involved. If it is the problem of automatic translation that the system is to solve, the set of the corresponding operations will depend primarily on the general formal model of the translation process underlying this system. One version of the model in question, proposed in [1] and discussed in more detail in [2], envisages the following main groups of operations:

1) analysis and interpretation of the initial text, simulating the process of perceiving and understanding its signification and denotation; ideally, it presupposes a semantic description of the text, as well as a model of the situation ("world" fragment) presented in it, being constructed from this text (possibly, via a number of intermediate representations);

2) translation proper, which is performed at a level R of some formal representation R_i of the initial text, derived from its analysis, and amounts to selecting translation equivalents for the units included in R_i: the result is an intermediate representation R_t of the target text, this representation being usually (although not necessarily) of the same level as R_i;

3) verification of the adequacy of the translation performed, by means of analyzing the resultant representation R_t and comparing the semantic description and the situational model obtained, with the semantic description and the model of the situation corresponding to the initial text;

4) generation (synthesis) of the target text by transforming the intermediate representation R_t formed during translation proper and assumed to be adequate by the verification procedure, into a sequence of actual word-forms and punctuation marks making up the target language text;

5) evaluation of the target text with a view to detect undesirable ambiguities and inaccuracies that might have slipped in during the synthesis process; it implies analyzing the text back to the R level and checking, whether the resulting representation R_t coincides with the representation R_t from which this text has been formed;

6) editing operations dictated by the checks and comparisons made: if the translation is judged to be inadequate they will consist in returning to the phase of translation proper and either substituting alternative translation equivalents for some of the previously selected ones, or reconsidering the entire procedure used and repeating it at a different ("deeper") representation level or in a different form (probably, resorting to synonymous transformations of the initial text at the R_i level); if it is the target text ambiguities and stylistic imperfections that are to be removed, better expressive means will be sought chiefly by actuating the system of synonymous transformations at the R_t level.

It is readily seen that the basic level of formal text representation from the standpoint of the above conception of the translation process is level R, directly concerned with the most important translation operations, primarily, the operations of translation proper, the scope of which is practically confined to the level in question, and the operations of synthesis ensuring the transition from the R-level representation of a text to its more "superficial" representations up to the text as such.

Some other of the operations mentioned involve also switching from the R-level to "deeper" levels of intermediate formal text representation and taking into consideration such supplementary factors as the essence of the situation described by the text to be translated, the semantic peculiarities of the vocabulary and the syntax of the two languages; the requirements of grammaticality and stylistic normativity (regularity) of the target text, and so on. The foregoing shows that these operations are mostly auxiliary in nature, their main purpose being to improve the content adequacy and the linguistic acceptability of the translation text formed through the use of the R-level representation; in a concrete automatic translation system based essentially on the formal model we have outlined, they may be reduced or even altogether omitted for various practical reasons.

However, whether these supplementary operations be included in an AT system or not, it is clear that the system will depend largely for its efficiency on the choice of the intermediate level R. It is precisely this basic level that we are now going to consider.

General Requirements

From the point of view of the purposes and peculiarities of the translation process, there are two opposite requirements that can be placed upon the intermediate formalized representation R in an automatic translation system.

On the one hand, insofar as translation boils down to transforming the surface structure of a text while preserving its content, it seems safe to assume that if some components of the text to be translated, some features of these components, or links between them are relevant for the content structure of this text, they may also prove of importance for choosing the correct translation equivalents for the text units. Consequently, the adequate representation R used in an AT system should be sufficiently "semantic" for all the necessary information concerning the components, links and features in question to be either explicitly given in this representation or, at least, to be easily obtainable from it. To put it differently, representation R of a text processed must reflect its semantic structure with sufficient precision and in sufficient detail.

On the other hand, the structures of the source and the target languages will, as often as not, have certain features in common, this leading to an inevitable neutralization of any analysis transformations involving such features, by the inverse transformations during the synthesis process. Such transformations will thus prove unnecessary for translation purposes, no matter how important they might be as regards the full semantic analysis of the text. Accordingly, representation R must be sufficiently "superficial" for its construction to incorporate the minimum possible of such superfluous transformations.

As we see, the second requirement provides a kind of limitation on the first one, restricting the extent and the methods of the explication necessitated by the latter, of the semantic structure of the text. Taking into account both of these requirements will most likely result in a kind of a compromise solution suggesting that information made explicit in representation R of a certain text should not include all the elements of its semantic structure; rather, it should cover only those of them which are apriori known to be extensively used in establishing inter-language correlations during translation.

With such a solution, however, one must be fully aware that real texts will contain a substantial proprtion of cases where some text information overlooked by our analysis might eventually turn out relevant for translation. If we do not want to give up the idea of adequately processing such texts as impracticable in principle, it seems useful to impose a third requirement on representation R - the requirement of "exhaustiveness" which may be formulated as follows. All information contained in a natural language text and not made explicit in its intermediate representation must be preserved within this representation; if possible, it should be preserved fully and without changing its original (natural language) form, so that there might be no accidental losses or distortions.

If so, the substitution of the formalized representation R for the original text will not exclude the possibility of

additional analysis amplifying the results of the standard analyzing procedure and providing access to some extra information that may be required. This is to say that the linguist describing the means of translating concrete language units within such a system will not be subject to the pressure of too stringent limitations originating from the conventions of the system, rather than from the nature of the material he deals with, and complicating his task (difficult enough as it is). Theoretically, he will be free to use any text information (both "superficial" and "deep")in any way he may find linguistically appropriate: whether as source units to be replaced by translation equivalents, or as conditions determining the equivalents chosen for some other units, or else as translation equivalents themselves.

The above principles are general enough to allow of various ways of implementing them in a concrete automatic translation project. We shall present here one attempt of such implementation made in defining the so-called Combined Structural Representation to be used in the system of Japanese-Russian automatic translation, now under development at the Institute of Oriental Studies of the Academy of Sciences of the USSR (Moscow)[3].

Combined Structural Representation (CSR)

Taking into account the typological correlation between the Japanese and the Russian languages, we consider it necessary to specify in the CSR of the initial Japanese texts, as well as of their Russian translations, five main aspects of text structure: the lexico-grammatical composition of the text processed, its predicate-argument organization on the semantico-syntactic level, the syntactic grouping of its units, the anaphoric relations between them, and the peculiarities of their linear arrangement. Within the CSR the corresponding five types of linguistic information about the text form separate components which will now be discussed in turn, mostly from the point of view of their consistency with the general requirements stated above.

Lexico-grammatical composition

The component of the CSR concerned with the lexico-grammatical composition of the text is intended to contain explicit descriptions of all lexemes present or implied (if ellipsis is the case) in the text under consideration, as well as of all grammatical (morpho-syntactical) elements accompanying them in the corresponding word forms or quasi-word forms (units taken to be functionally analogous to word forms). The descriptions required must include, apart from the symbols of the units involved, information about their meanings within the text in question and about their relevancy or irrelevancy as regards the process of its translation.

The operations necessary to obtain this component of the CSR when analyzing the initial Japanese text will evidently comprise isolating separate word forms and determining their internal structure (in terms of lexemes and morphologic markers), resolving ambiguities for all units established; eliminating synonymy where it is manifested as supplementary distribution or free variation of morphologic units; detecting phraseological word combinations and reducing them to a one-word symbol; giving special labels to those word forms or parts of word forms which play an auxiliary role in the text analyzed and require no special translation equivalents; filling in the units omitted in the source text if their absence obscures its structure and hinders the translation process (due to the differences between the rules of linguistic ellipsis in the two languages), etc.

From this it follows that the lexico-grammatical composition of a text cannot be definitively established in the course of its analysis without drawing upon information about its structural characteristics. The same kind of information is also needed when working with this component of the CSR in the synthesis process (chiefly in connection with such means of expressing structural relations as grammatical agreement and government, typical of the Russian language).

Therefore, in deciding what language units are to be described as permissible in the given component of the CSR, and what status is to be attributed to them within its framework, specifically, which units it is best to treat as individual words and which ones should rather be regarded as meaningful parts of words - morphemes (the problem being of particular importance for Japanese where no regular graphical means are used in writing to separate words from each other), we believe it advisable to pay special attention to the functions of the corresponding units in the general structure of the text and in the system of operations used for its processing. With this aim in view, we have devised an operational criterion of distinguishing words and their meaningful parts, based on the principle of the homogeneity of the levels of text processing[4] and on the requirement that each level's units should have structurally significant functions within the level

itself, while there should also exist a well-defined (although not necessarily one-to-one) correspondence between certain subsets of units belonging to the adjacent levels of processing. According to this criterion, the status of separate words is justified, among others, for such Japanese units as the so-called "causative voice" marker -seru/-saseru, the "conditional mood" marker -ba, the negation marker -nai (at least, in conditional contexts) and some others. Among units functionally analogous to independent word forms (and, consequently, appearing as such within the CSR), are also classified punctuation marks which are, to our mind, quite similar to words in that they can be meaningful and can correspond to definite translation equivalents (or play the role of such, cf. Japanese ka vs. Russian ?).

In this way, so far as the position of a unit in text structure and in the system of translation transformations is related to the meaning of this unit, our general principles of describing the lexico-grammatical composition of texts in their CSR conform to the requirement of its "semanticity". On the other hand, the "exhaustiveness" requirement is also met, since we make it a point not to leave out of the CSR any text elements, up to those that serve essentially as surface markers of other linguistic units made explicit in this representation, and do not themselves participate to any significant extent in the semantic operations provided in the system (e.g. Japanese "case" particles; Russian morphological categories of case, gender and number of adjectives; "surface" linguistic expression of "lexical functions" and their translation equivalents, etc.).

Predicate-argument organization of the text on the semantico-syntactic level

This component of the CSR represents semantico-syntactic links between words and/or quasi-words corresponding to their predicate-argument relations and, accordingly, constituting meaningful text units. It is common knowledge that the surface expression of these units is language-specific while their semantic content is generally assumed to be of a more or less universal nature. So in translation they must either remain essentially the same (naturally, with all the necessary modifications of their surface markers) or must be transformed by certain formal rules depending on the semantic interpretation of the links in question and on their relation with the meaning of the units linked.

The lexico-syntactic translation transformations mentioned are most commonly used where the source and the target languages have appreciable typological differences. This is precisely the case with the Japanese-Russian correlation (a simple example: karo-wa mannenhitsu-o nusumaremashita, lit. "he was stolen a pen", transl. У него украли ручку "he had his pen stolen"). Bearing this in mind we have chosen the dependency grammar to represent the predicate-argument structure of texts in their CSR, preferring it to its alternative - the immediate constituent system, for according to a number of specialists, this type of transformations is easier to describe in dependency terms.

One of the central linguistic problems connected with presenting the predicate-argument structure of a text in its CSR is which among the various (and often semantically overlapping) dependencies between the text units should be selected for explicit description. In solving this problem we proceed from the principle of the possibility of "immediate semantic substantiation" of the dependencies to be selected. It can be specified as the following requirement bearing on the ways and methods of describing words and grammatical constructions when compiling the linguistic information for the automatic translation system:

- all syntactic dependencies registered in the CSR of a certain text must realize some semantico-syntactic valencies of the lexical or grammatical units present in it (and usually forming part of the lexico-grammatical composition of the word forms or quasi-word forms linked by the corresponding dependencies).

These valencies, in their turn, must directly correlate with the semantic characteristics of the units they are ascribed to, semantic considerations viewed as the major factor underlying their assignment to those units. One important consideration of this kind consists in preferring the descriptions where the maximum possible of the valencies envisaged could be realized in concrete texts by two-word combinations and the maximum possible of such combinations could be checked for their semantic acceptability (consistency) without regard to any units outside them.

Apart from the situations where some of the syntactically linked units perform in the text processed auxiliary functions (thus having no independent semantic content) the application of the above criteria can only be limited for reasons of economy and effective controllability of

the linguistic description.

From the above it can be inferred that the linguistic information used to reveal and/or process the predicate-argument structure of concrete texts should combine data on the means of surface expression of the links involved (i.e. word order, function words, etc.) with fairly detailed semantic descriptions of the words to be linked and of their combinatorial potentialities. To provide the formal tools necessary for constructing such descriptions we have devised a special formalized semantic language SL[5], the characteristic properties of which can be briefly outlined as follows.

The vocabulary of SL comprises three categories of the so-called semantic elements: categorial elements, encyclopaedic elements and identifying elements. Among these the leading role belongs to the categorial elements which are given special descriptions constituting a kind of formalized semantic grammar of the natural language. The syntax of SL, used to combine semantic elements into semantic formulae, accounts both for the semantic relations established between the components of such a formula and for its communicative organization determining the behaviour of its components as regards the logic operations that can be applied to the formula as a whole. From the formal point of view a semantic formula is a linear sequence of symbols, structurally equivalent to a special type of a dependency tree where the nodes can be labeled by the symbols not only of single semantic elements, but also of their combinations(subtrees)of any length.

Semantic formulae can be employed to express: 1) semantic definitions of natural language units (from a separate word up to a whole text); 2) their paradigmatic semantic features; 3) their syntagmatic semantic properties (semantic interpretations of their syntactic valencies).

An important distinguishing characteristics of SL is that it affords formal derivability of information about the semantic paradigmatic and syntagmatic features of language units from their semantic definitions. This helps to make the semantic descriptions of these units more compact (by eliminating the unnecessary reiteration of essentially the same data) and to improve their reliability, owing to the possibility of more objectively evaluating the adequacy of semantic definitions on the basis of such a criterion as the degree of correlation between the syntagmatic properties of a unit derivable from its definition, on the one hand, and its actual semantic combinability as observed in real texts, on the other hand. Moreover, it increases the range of linguistic facts explainable on semantic grounds. Thus, it becomes possible to give uniform rules (unattainable if one stays within the bounds of purely lexico-syntactic phenomena) for the selection of the correct morpho-syntactical markers (as well as for the appropriate synonymous transformations and logical deductions - operations commonly used as translation devices) when handling constructions with such Russian verbs as грозить ("run the risk"), опасаться ("fear"), ожидать ("expect"), успевать ("be in time"), etc., taking predicate words as their arguments. These rules will enable us, for example, to choose the correct Russian sentence

К раненому опоздали с помощью

("Help came late to the wounded man"), rather than

* Раненый опоздал с помощью

("The wounded man was late with help") as translation of the Japanese sentence

Keganin-wa teate-ga okurete shimatta.

With semantic definitions of words formulated in the SL terms, all syntactic dependencies linking these words in texts can be interpreted (for the most part, unambiguously) as semantic relations between certain elements within their definitions, and replacing a word by its semantic definition will not alter the general form of the predicate-argument structure of the text. The effect is that in the framework of the predicate-argument component of the CSR the contradiction between the "semanticity" and the "superficiality" required of it, turns out to be to a large extent eliminated. For one thing, any fragment of the predicate-argument structure of a text can be interpreted (developed) as a structure of semantic elements and relations; for another, the scope of such interpretation does not depend on any but linguistic considerations, and if no transformations affecting the internal semantic structure of words or relations between them are necessary for translating a certain text fragment, the latter need not be semantically interpreted, no matter whether this kind of interpretation be indispensable for some other fragments of the same text.

Syntactic grouping of text units

This type of structural information about the text concerns the grouping of the words contained in it into larger combinations possessing certain syntactic and/or semantic independence, which makes it advisable to treat them as separate units at least at some stages of processing the text in question. In a way such

information is analogous to the information about the constituent structure of the text. The difference is, though, that the aspects of syntactic word-grouping included in the CSR of a text are limited to those that carry semantically relevant information lacking in its dependency structure[6] (and, for that matter, not always directly expressible in the classical constituent marker form, either).

For the present, the given component of the CSR of a text is supposed to specify only the word groups established within connected fragments of its dependency tree in situations where the composition of such groups and their boundaries are important for some of the operations employed to process it, such as ascertaining the domain of the quantifiers; distinguishing between descriptive and restrictive attributes; revealing the full form of some types of elliptical constructions (e.g. those with co-ordinative reduction); deciding on whether it would be safe to employ transformations disjoining elements of some word-combinations within the text's dependency structure or linear representation (it seems reasonable to mark the combinations excluding this kind of lexico-syntactic transformations as a special type of syntyctic word-groups), etc.

The relevancy of the data on syntactic word-grouping for translation purposes can be illustrated by the Japanese sentence

Watakushitachi-no tsukau nichi-yōhin-de nagai aida tsukatte mo hera-nai mono-wa nai,

meaning "Among the things we use daily there are none that could be used for a long time and still remain as good as new".

If the data in question is not taken into account here we are liable to distort the presuppositional structure of the sentence by giving it the "literal" translation:

*Среди используемых нами вещей домашнего обихода нет таких, которые бы не изнашивались, даже если ими пользоваться долгое время

("Among the things we use daily there are none that do not wear out, even if used for a long time"),

having the evidently false implication that the longer things are used the less they wear out (cf.: Нет вещей, которые бы не изнашивались, даже если ими пользоваться очень аккуратно"There are no things that would not wear out even if they are taken good care of").

The origin of this undesirable implication can be explained two-fold. The first explanation is that one of the word-group boundaries in the given Russian sentence separates the negation не ("not") from the whole of the fragment following it in the linear sequence of this sentence:изнашивались бы, даже если ими пользоваться долгое время ("wear out even if they are used for a long time"), so that the fragment cited is interpreted as an integral semantico-syntactic unit, this giving rise to the implication to be avoided. According to the other explanation, the boundary responsible for the interpretation of the Russian sentence runs between the whole of its initial fragment Среди используемых нами вещей домашнего обихода нет таких, которые бы не изнашивались ("Among the things we use daily there are none that do not wear out") and the remaining sequence даже если ими пользоваться долгое время ("even if they are used for a long time"). From this standpoint, the false implication is accounted for by the possibility, suggested by grouping the sentence units into the above two fragments, of interpreting and/or transforming these independently of each other, thus obtaining

Любые из используемых нами вещей домашнего обихода изнашиваются, даже если ими пользоваться долгое время ("All of the things we use daily wear out, even if used for a long time").

No matter which one of the two explanations be taken as true (the second one seeming more plausible, while the first one suggesting simpler check-ups in processing texts) it is clear that the translation problem is to achieve in Russian the same syntactic grouping as in the original, by introducing the corresponding lexical and/or positional (linear) modifications, e.g.:

Среди используемых нами вещей домашнего обихода нет таких, которые бы даже при длительном пользовании оставались неизношенными.

Another (and, probably, more ordinary)case of using data on syntactic word-grouping in translation can be exemplified by the sentence:

Rōdōsha-ga (tsuyoku danketsu-shite seiji-teki yokyu-o dasa) nakereba warewa-re-no seikatsu suijun-o itsumade-mo yoku saseru koto-ga dekinai.

Here it is essential that the negation marker, as well as the expression of condition, which in the translation sentence must take a position different from the one its Japanese counterpart occupies in the original word-sequence, should not interpose between the two members of the co-ordinative-type word-group present in the sentence (for clarity, we have enclosed this group in brackets). That is, the translation must be (English being structurally similar to Russian in this

respect):

If the workers do not (unite and put forward political demands) we shall never be able to raise our life level

and not

*If the workers unite and do not put forward political demands..."

Generally speaking, the correct translation of the last example (as well as of other constructions explainable in terms of co-ordinative reduction) could also be obtained without recourse to the information about syntactic word-grouping. Instead, one could use a "deeper" description of the text to be translated, with elliptical constructions transformed into their full representations. However, this kind of transformation would be basically superfluous, for in the synthesis process it would be necessary to reduce the constructions in question back to their elliptical form using but slightly different rules. It seems therefore preferable for the operations of translation proper to result directly in an elliptical construction analogous to the original one and differing only in details of its surface expression (such as the position of negation in the above example), specified by the subsequent synthesis procedure.

So we see that while the component of the CSR under discussion registers only semantically significant phenomena of text structure, the means of representing them in it remain essentially superficial, so as to satisfy both the "semanticity" and the "superficiality" requirements.

Anaphoric relations between text units

For interpreting texts in respect of their signification and especially denotation, the structure of anaphoric relations between their units is on the whole no less important than their predicate-argument structure. However, the anaphoric structure is expressed mainly by lexical repetition, and this can be easily accounted for if we require that as long as one text is dealt with, one and the same translation equivalent should be selected, so far as possible, for all occurrences of one and the same lexeme (lexeme being defined as a word taken in one of its various lexical meanings). Given this requirement (which appears to be natural enough and, but for some special cases, easy to comply with), there is no need to include this structure in the CSR in its full form. It seems sufficient to indicate it only for those types of language units which directly depend for their translation on the properties of their antecedents in the text at hand.

In Japanese (as also in other languages) there are two types of such units.

The first type are pronouns: when translating, say, the pronoun sore, the choice of one of the words: this, he, she, it, they, one, etc.,- as its text equivalent will be determined, among other things, by the syntactic class of the unit chosen as the equivalent of its antecedent. If this unit is a noun, one will also need to know its number and (for Russian) gender.

The second type of units which cannot be translated properly without information about their antecedents is more specific. These are words which are graphically identical with components of more complex units, also lexicalized from the point of view of their semantic behaviour, and which can function as structural substitutes for the latter. When used in this function, such words must be replaced either by the translation equivalents of their antecedents, or by pronouns (with the data on these antecedents used in the same fashion as in translating usual pronouns). Anyway, their own translation equivalents are ruled out.

Thus, the word nimotsu, meaning "luggage" if used independently, will be translated as "them" or "these parcels" in the context of the sentence

Konimotsu-gakari-ga mazu nimotsu-no megata-o hakarimasu,

where nimotsu is substituted for konimotsu ("parcel"):

The clerk dealing with parcels first weighs them (these parcels).

As regards all other types of lexical units, our approach is that the existence of anaphoric relations between them should be checked and the relations themselves registered in the CSR for further processing only in those infrequent situations (due mostly to dissimilarities in the combinatorial properties of the original language words and of their translation equivalents, this necessitating the use of synonymous transformations) where it is impossible to fulfil the above requirement of translating different occurrences of the same lexeme by the same equivalent, and one has to make sure that employing different equivalents in this case does not affect the original anaphoric structure of the translated text.

Linear arrangement of text units

In dealing with linear arrangement of units in a text in the framework of an automatic translation system, it is important to distinguish between two types of their positional (word-order)

relations requiring different processing during translation.

If the first type of such relations occurs between two text units, the position of one of them in respect to the other is merely a surface syntactic marker showing the presence (or absence) of, say, some semantico-syntactic link between the two, an anaphoric relation between them, a syntactic word-group boundary, and so on. In case of the second type such position is meaningul in itself, irrespective of whether it should or should not be taken into account when establishing certain syntactic links or boundaries: it shows the relative positions of the units in question in the communicative structure of the text (i.e. from the point of view of its functional perspective).

It should be noted that the opposition of these two types of positional relations is not the same as that of rigid (fixed) and free word order: while free word order is always "semantic" to some extent, rigid word order can, to our mind, correspond to both cases, depending on whether the given syntactic construction with rigid word order correlates in the language under consideration with any alternative constructions providing the same predicate-argument structure and/or syntactic grouping of their components, but assigning them a different linear arrangement (a possible example of such alternative constructions which can be considered as dependent for their selection on the word order required, rather than vice versa, is furnished by predicative constructions differing in their voice value).

Guided by the "exhaustiveness" principle, we judge it expedient for the CSR to contain information both about the "meaningful" and the "auxiliary" type of word-order relations, though represented and employed in different ways.

The sphere of employment of the "auxiliary word-order information is practically limited to the analysis and synthesis procedures. During the analysis phase this information serves mainly as a means of revealing and formally representing units and constructions pertaining to other components of the CSR; in the synthesis phase it is used to obtain the correct form of the same type of units and constructions in the target language. The corresponding facts of the linear arrangement of the text do not play any independent role either in its semantic processing or in choosing translation equivalents for its units, so it is perfectly sufficient to regard them as just one of the various features of the units and constructions involved, important enough to be registered in their linguistic descriptions, but constituting no separate objects of description. To incorporate these facts in the CSR, we resort to numbering the words in the text processed in the order of their successive occurrence (the resulting numbers used also, in combination with some other data, as their identifiers throughout the processing).

If, on the contrary, a construction is characterized by a meaningful word-order relation between its lexical components, it is given the status of a special "positional unit", distinct from the construction itself and represented explicitly in the CSR. Such a unit directly participates in semantic operations, including those of translation proper, which means that it must have its own description (in particular, its own translation equivalent). It stands to reason that the range of inter-language correspondences involving positional units of either the source or the target language is not restricted to this class of units alone, as the communicative organization of text can also be conveyed by some types of syntactic constructions and lexical elements. An example is the Japanese particle ga as used in independent sentences (or, sometimes, in the main clauses of complex sentences), where its best Russian equivalent (if the same type of predicative construction is used) is the reverse order of the subject and the predicate.

As we see, here also, as in the other components of the CSR, there is a compromise between the "semanticity" and the "superficiality" requirements. On the one hand, explicit indication of the word-order relations found to be meaningful in the text processed characterizes some aspects of its semantic structure. On the other hand, the form of "positional units" chosen to represent them is rather superficial in that it does not display the semantic correlations underlying the interchangeability of these units with other structural text features(such as the selection of the nexus vs. junction form of expressing the predicate-argument dependencies between text units; the use of "relational" words, of the Operi or Funci type and the like; the occurrence of emphatic particles and constructions, etc.).

Conclusion

In the foregoing we have tried to show the way the Combined Structural Representation of text reflects the requirements of "semanticity", "superficiality" and "exhaustiveness" formulated at

the beginning of the paper as essential for the basic level of formally representing text structure in an automatic translation system. We shall now briefly recapitulate the points.

The "semanticity" requirement is accounted for in the CSR, in the first place, by the very possibility provided in it of explicitly describing the five most important aspects of text structure and composition, as stated above. The quest for "semanticity" forms also the basis of the principles we employ in selecting concrete information to be made explicit. Among these one can mention the criterion of structural significance of the units to be represented in the CSR as separate words or quasi-words; the principle of "immediate semantic substantiation" of the predicate-argument syntactic relations registered in it; the requirement of supplying the elements of the lexico-grammatical composition of the text under consideration, as well as of its linear arrangement, with indications of their meaningful or auxiliary role within this text; the employment of a special formal language to define the semantic properties of words and word-combinations, etc.

The "superficiality" of the CSR is seen, among other things, in the fact that this level of text representation envisages the use of lexico-syntactic translation equivalents and does not necessarily require decomposition of lexemes into combinations of smaller units of meaning, such decomposition considered appropriate but in comparatively rare cases of descriptive and interpretative translation. Other features of the CSR originating from the "superficiality" principle are absence of exhaustive information about the anaphoric structure of the text, inclusion of only those data on syntactic word-grouping which are of importance for the translation process, direct translation of elliptical constructions, wherever possible, etc.

Finally, the "exhaustiveness" requirement is specified as what may be called the "lose-nothing" principle of constructing the CSR. It means that when special labels are formed in it to explicitly display various structural elements implicitly present in the surface form of the text at hand, the surface text markers (such as the "auxiliary" type word order; morphological features expressing grammatical agreement or government; function words and punctuation marks having no independent translation equivalents, and so on), though having been used already to reveal those elements, are not eliminated from the representation being formed. They are merely supplemented by the designations of the elements revealed, as well as by formal indications of their own auxiliary nature, and thus remain accessible for any further analysis that might prove useful, should it turn out that their functions in the text are not limited to just identifying the units already made explicit.

Notes

[1] З.М.Шаляпина. К проблеме построения формальной модели процесса перевода. - В кн.: Теория перевода и научные основы подготовки переводчиков. Часть II. М., 1975, с. 165-172.

[2] Z.M.Shalyapina. Automatic translation as a model of the human translation activity. - International Forum on Information and Documentation, 1980, vol.5, No.2, p.18-23.

[3] An earlier version of text representation aimed at incorporating the principles proposed had been developed in the framework of an Anglo-Russian automatic translation project and described briefly in: З.М.Шаляпина. Англо-русский многоаспектный автоматический словарь (АРМАС). - Машинный перевод и прикладная лингвистика. Вып. 17. М., 1974, с. 7-67.

[4] The notion of levels of text processing is not identical with the notion of levels of text representation (although there certainly exist some strong correlations). Linguistically, the former corresponds rather to the notion of language tiers introduced in: И.Ф.Вардуль. Основы описательной лингвистики. М., "Наука", 1977.

[5] A detailed formal definition of this language and a description of some of its linguistic interpretations are given in: З.М.Шаляпина. Формальный язык для записи толкований слов и словосочетаний. - Проблемы кибернетики. Вып. 36. М., 1979, с. 247-278.

[6] There is also a paper on a French-Russian automatic translation project where a similar type of structural information is mentioned as necessary (see: Ю.Д.Апресян и др. Лингвистическое обеспечение в системе автоматического перевода третьего поколения. М.,1978,с.13). In our case of Japanese-Russian translation, however, such information seems to require more attention due to wider differences between positional and other rules of expressing the corresponding constructions in the two languages.