Grammar in the Light of Machine Translation

PAUL L. GARV1N

Department of Linguistics,
State University of New York
at Buffalo.
C106 Spaulding Quadrangle.
Buffalo. New York, 14261 USA

Units and relations applied in conventional language description often turn out to be inappropriate for the purposes of formalized description, in particular, when developing a machine translation system. Word classes and categories of Russian grammar are outlined as used in parsing according to the FULCRUM approach. Subdivision into classes is not identical with classification of words by parts of speech. Content words are subdivided into nominals (conventional nouns and nominal pronouns), attributives (adjectives, long-form participles, adjectival pronouns, numerals), predicatives (finite verb forms, short-form adjectives and participles, and predicatives proper like HBAban), infinitives, adverbs, and gerunds. Word classes of nominals, predicatives, and attributives are additionally considered in more detail.

This paper deals with the conception of grammar, more particularly, the Russian grammar developed for the FULCRUM machine translation project by the author and his associates some years ago*.

Most workers in the field agree that grammar enters into the MT process in two systems components: (1) the codes contained in the machine dictionary; (2) the recognition rules which allow the system to process the input text in order to arrive at the sentence image representing its syntactic (and other) structure. It is now generally accepted that a correct parse of the sentence is the prerequisite for all subsequent processing, be it for purposes of translation or some other information handling.

This paper is concerned with the coding system employed in the FULCRUM approach, the view of Russian grammar that it represents, and the reasons for this view of grammar which, in a number of respects, differs from the usual one.

The role of the coding is to reflect the grammatical and other linguistic potential of each dictionary entry. Given the crucial role of parsing, one of the primary tasks of the codes is to reflect the syntactic function potential of the dictionary entries. This is the aspect of coding that will be of primary concern here.

The word classes and grammatical categories of conventional Russian grammar had to be re-evaluated in the light of machine translation: since many of these grammatical characteristics traditionally (and also in some modern work) are, at least in part, assigned on the basis of morphological identity, their status had to be reviewed in the light of their primary syntactic function (which is decisive for machine translation purposes, while morphological composition is much less relevant). This has led to a reorganization of the conventional view. The most conspicuous departures from this view are in the area of word classes and subclasses — more specifically, the classes and subclasses of 'content words', since for 'function words' there is no strong tradition of classification on morphological grounds. This is what will be emphasized here.

Thus, for instance, while most grammars recognize the multiple word class status of participles, their primary assignment is on the basis of their morphological stem identity and part of their affixation pattern as a subcategory of verbs. Taking the case of 'long-form' participles, however, it is clear that their primary syntactic role is that of modifying nominal structures. For machine translation purposes, the FULCRUM approach recognizes this by including them in the same major word class with conventional adjectives, that of attributives. Similarly, most grammars include 'short-form' adjectives among adjectives as special morphological forms of these. In the FULCRUM grammar code, these are included, together with finite verb forms, in a word class of predicatives of which the former are subclasses (nonfinite verb forms such as infinitives and gerunds, and participles— see above—are then assigned to other word classes).

In the following, first a survey of the word classes and subclasses established for the FULCRUM project will be given, together with a brief discussion of their justification in terms of syntactic functioning. Then a more detailed sketch of the word classes of predicatives and attributives will be given; and finally, some comments on the subcategorization of nominals, more particularly of nouns, will be presented.

The following are the word classes of content words established for the FULCRUM system: nominals, attributives, adverbs, predicatives, infinitives, gerunds. They will now be briefly characterized.

Nominals. These include all of the nouns and nominal pronouns of conventional grammar: the syntactic function potential in terms of which they are defined is their ability to serve as heads of, or constitute, nominal

* The MT Project at Wayne State University, conducted at »bout the same time, also used most of the principles of the FULCRUM approach.

blocks (the machine translation analog of noun phrases). This word class will be discussed in more detail below.

Attributives. These include the adjectives, 'long-form' participles, adjectival pronouns and numerals of conventional grammar; the syntactic function potential in terms of which they are defined is their ability to serve as modifiers to nominals. As already noted, they are further subdivided into a nongoverning and governing subclasses; this word class will likewise be discussed in more detail below.

Adverbs. These include the adverbs (including comparatives of adverbs) of traditional grammar.

Predicatives. These include the finite verb forms, 'short-form' adjectives (including comparatives of adjectives) and 'short-form' participles of conventional grammar; the syntactic function potential in terms of which they are defined is their ability to serve as heads of, or constitute, predicate blocks (i. e, constructions forming predicates in the narrow sense, not including objects or 'adverbials'). This word class will also be discussed in more detail below.

Infinitives. These include the infinitives of conventional grammar; the syntactic function potential in terms of which they are defined is their ability (1) to form part of predicate blocks (see above) when governed by predicatives; (2) to constitute, or form part of, arguments (subjects or objects) of clauses.

Gerunds. These include the gerunds of conventional grammar; the syntactic function potential in terms of which they are defined is their ability to be heads of, or constitute, predicate blocks. They differ from predicatives by the fact that the latter are also capable of taking subjects, while gerunds are not. The predicative function of gerunds is thus more limited than that of predicatives.

The way in which the word class information is utilized by the FULCRUM system is the following: the grammar code of each word is 'read' by the processing algorithm in the appropriate place of the flow of analysis, and the word class and associated information triggers an appropriate subroutine that operates on the code for purposes of syntactic recognition. As the subroutine goes into effect, it computes the function that the word in question concretely has in the particular textual passage in which it is contained (to the extent allowable at that point in the processing).

In the next sections, a more detailed discussion of predicatives, attributives and nominals will be given, showing the application of the just mentioned general flow to these word classes. For each of these, the following characteristics will be discussed in varying degrees of detail: subclasses and subcategorization, additional grammar coding such as agreement code and government code, and the searches that are triggered by these. The word classes to be discussed have been selected because they are believed to aptly illustrate some of the insights into grammatical phenomena that FULCRUM-oriented research in machine translation has provided.

Detailed discussion of predicatives. As already noted, these are defined by their ability to serve as heads of, or constitute, predicate blocks. They share this ability with gerunds, but unlike the latter, they are also capable of taking subjects; they share with infinitives the ability to be included in predicate blocks (but not the ability to constitute such blocks in their entirety), but unlike the latter, they cannot form part of, or constitute, arguments of clauses.

Thus, whenever the FULCRUM algorithm 'reads' a predicative grammar code, it will be 'alerted' to the possibility of a predicate block which in turn usually may be assumed to serve as the fulcrum of a clause. By reading the additional grammar coding of the predicate block (primarily the predicative, but also other accompanying and dependent words) searches for other constituents of the clause, such as its subject and object (s) will be triggered.

Predicatives fall into subclasses on the basis of the kind of agreement pattern they exhibit: this information replaces the conventional word class differences that were suppressed when the FULCRUM-based word class of predicatives was established. These will be discussed after the role of the agreement and government codes has been pointed out.

The importance of 'the agreement codes of predicatives lies in the fact that it is in terms of these codes that potential subjects can be searched for, since — as is well known —the predicate must agree with its subject Thus, given a predicative with a 'feminine singular' agreement code (such as неприятна), the subject of the predicate constituted — entirely or partially — by this predicative will be a feminine singular nominal block (such as проблема). Government codes, on the other hand, indicate the kind of objects that are to be searched for, since —as is likewise well known — predicates govern their objects. Thus, given a predicative with an 'instrumental' government code (such as управляет), the object of the predicate constituted — entirely or partially— by this predicative will be a construction in the instrumental (such as фабрикой).

The way in which the agreement pattern serves as a criterion for the subclassification of predicatives is shown by the following oppositions.

The first opposition concerns the possibility of being differentiated for number and/or gender; a given predicative may or may not be differentiated in this respect. Predicatives not so differentiated are illustrated by the conventional 'short-form' comparative adjectives (such as сильнее); these can agree with subjects showing any number and gender. Predicatives that exhibit differentiation in turn fall into two subclasses: (1) Those that are differentiated for number only—these include the nonpast finite verb forms of conventional grammar (such as проходит or говорят); their subjects can exhibit only one of the two numbers but any of the three genders (e. g., едет can have a subject only in the singular, but of any gender —such as поезд, бригада or животное; идут can have a subject only in the plural, but of any gender — such as работы, разговоры, занятия). (2) Those that are differentiated for both number and gender; these include the past-tense verb forms and 'short-form' adjectives and participles of conventional grammar (such as пошел, важна, ясно or даны); their subjects, as is well known, can exhibit only the particular number and gender of the predicative.

The second opposition concerns impersonality: a given predicative cannot, can or must be impersonal (the first is exemplified by сидит, the second by кажется, and the third by надо). A predicative of the third subclass will not normally take a subject (although it may take a dative of reference, see below), one of the second subclass may or may not, one of the first subclass will.

The third opposition concerns genitives or datives of reference: a given predicative cannot, can, or must take a genitive or dative of reference. Genitives or datives of reference take the place of subjects in certain 'constructions (some linguists will consider them representatives of 'deep' subjects). Thus, most predicatives take only subjects in the nominative and will carry the grammar code 'cannot take a genitive or dative of reference'. A predicative such as 6fauio, when modified by the negative particle, will acquire the notation 'may take a genitive of reference' since it can be construed with either the usual nominative subject (условие не было ...) or a genitive of reference (его не было), while a predicative such as нет will take the notation 'must take a genitive of reference'. A predicative such as нужно will take the notation 'may take a dative of reference' since it can be construed with either the usual nominative subject (занятие нужно) or a dative of reference (нам нужно...). Finally, predicatives such as надо or хочется will take the notation 'must take a dative of reference'.

Detailed discussion of attributives. As already noted, these are defined by their ability to serve as modifiers to nominals. The subclassification into nongoverning and governing modifiers is based on a crucial difference in syntactic function potential: governing attributives can govern clausal-argument-like dependent structures, nongoverning attributives cannot. By governed structures, we mean structures 'strongly' governed by the given attributive, not structures that can be 'weakly' governed by any attributives other than adjectival pronouns or numerals. Thus, any such attributive can 'weakly' govern a structure of the sort по сравнению с..., as in важный по сравнению с.... On the other hand, attributives such as являющийся are considered governing in our sense since they 'strongly' govern the dependent structure, in this case, an instrumental object as in являющийся зультаом. Another interesting aspect of attributives is whether or not they are modifiable by самый: болшой is, второй is not.

Detailed discussion of nominals. The most interesting aspect of this word class is the kind of 'strict subcategorization' that applies to it. Conventional Russian grammar correctly recognizes a morphologically marked distinction between animate and inanimate nouns, as reflected in the use of case endings. There is another syntactic/semantic difference between animate and inanimate nouns which is not much stressed in most grammars: the kinds of predicates for which nominals of the two subcategories can serve as subjects. Predicates such as заявил, решил tend to have animate nominals as subjects (исследователь заявил, руководитель решил). However, a number of nouns which are morphologically inanimate may also serve as subjects of such 'animate' predicates; in fact, they occur in this role quite commonly (комиет заявил, совет решил, etc.). In the FULCRUM grammar code, a separate subcategory status has been established for such nominals, that of pseudo-animates,

FULCRUM-oriented research in machine translation has shown what kinds of insights result from the overlap of system-design and linguistic considerations in the process of developing the design of a machine-translation system of this kind. In addition to new discoveries such as the revelation of hitherto poorly researched details of the grammar (e. g., the subcategory of pseudo-animates just discussed), it involves a great deal of re-examination and re-emphasis in areas of the structure that are generally well known but are not usually presented in a manner most relevant to the needs of machine translation as seen from the FULCRUM pers-, pective.