THE RUMORS SYSTEM OF RUSSIAN SYNTHESIS

Max I. Kanovich, Zoya M. Shalyapina

Institute of Oriental Studies, Russian Academy of Sciences,
Rozhdestvenka str., 12, 103753 Moscow, Russia

Abstract

The RUMORS synthesizer of Russian is an integral part of the JaRAP experimental system of Japanese-Russian automatic translation, although it can also have other applications. Morphologically, it is based, primarily, on A.A.Zaliznyak's model of Russian inflexion. Syntactical functions of RUMORS rely on word-order and dependency data as input information. The synthesizer is implemented on IBM PC, MS DOS, in Turbo Pascal.

1 General information

The RUMORS system of RUssian MORphological and Syntactical synthesis has been developed as part of the JaRAP experimental system of Japanese-Russian automatic translation, described in (Modina,Shalyapina,1994). Its operation is, however, completely independent of the other components of the JaRAP system, so that RUMORS could be used in any other AT system irrespective of its source language. Basically, RUMORS constitutes a system in its own right, which can also be used for purposes other than translation, e.g., as a computerized reference book of Russian morphology and the simplest phenomena of syntactic government and agreement (for students and teachers of Russian), as part of a spell-checking system, etc.

The RUMORS synthesizer has two major modes of operation: the QUERY mode and the TASK mode.

In the QUERY mode of operation, RUMORS accepts, as its input, a separate syntactico-morphological query (entered from the keyboard) which represents the lexeme to be processed and the syntactical and morphological characteristics specifying the word-form to be obtained by the processing. The output is, primarily, the desired word-form of the input lexeme or, if necessary, a *periphrastic substitute* for this word-form.

After obtaining this output, the user can switch at will to the FULL PARADIGM submode of the QUERY mode. In this submode, RUMORS generates all synthetic word-forms of the input lexeme (or of the last lexeme processed while obtaining a periphrastic word-combination). It also offers a set of menus allowing the user to modify his initial query by choosing additional morphological categories from these menus.

In the TASK mode, the input data is a sequence of queries fed from the special TASK file. Apart from lexical, morphological, and syntactical information contained in each query, the TASK mode of operation makes considerable use of word order data which is essential, among other things, for processing prepositional, adjectival and noun phrases. The output is the sequence of word-forms manifesting the phrases or sentences specified by the input sequence of queries.

Both in the QUERY and in the TASK modes, the output is displayed on the screen and written simultaneously in the special SOLVE file. If required by the user, it may also include the alterations made in the queries processed and the database information used in their processing. Inasmuch as the simulation by RUMORS of the linguistic processes involved in Russian synthesis is faithful enough, this auxiliary data could be valuable by itself (e.g., for learning or teaching Russian), aside from its significance for debugging and controlling purposes.

2 Synthesis functions envisaged by RUMORS

2.1 Morphological functions

The morphological functions of RUMORS cover all aspects of Russian inflexion, as well as some semantically basic lexico-morphological relationships.

The **inflexional functions** are initiated after the input query has already been subjected to syntactical and lexico-morphological operations, which may have modified its initial form. At this stage, it contains nothing but the lexeme and the inflexional categories specifying its desired word-form. If some categories needed for complete specification of this word-form are not explicitly stated in the query, they are settled by default. E.g., a query containing nothing but the lexeme of a verb is taken to describe the finite

form, indicative mood, present tense, active voice, 3d person singular of this verb, so that the query, say, *брать* produces the form *берет.*

The inflexional model of Russian implemented by RUMORS is the one proposed and detailed by prof. A.A.Zaliznyak (1977). Its important virtue is that the generation procedures it envisages represent those to be expected of human speakers of Russian more faithfully than any other known model, while the requisite database information is very compact.

Our version of Zaliznyak's inflexional model differs from its description in (Zaliznyak,1977) in two respects. On the one hand, we have reduced the scope of the original Zaliznyak's model, implementing it only in so far as written Russian is concerned. As a result, quite a number of the particulars of Russian accentuation registered in (Zaliznyak,1977), namely, all those that are relevant for oral speech only, have been ignored.

On the other hand, we have extended the model to cover analytical word-forms. Moreover, we have introduced a new type of morphological functions, the *periphrastic functions* allowing RUMORS to produce output that makes sense even if the required word-form is non-existent (e.g., due to the lexeme having a defective paradigm or to the combination of categories in the query being beyond the scope of Russian inflexional morphology). E.g., the future tense 1st person singular of the verb

победить 'win'

(which does not have this form) is paraphrased as

смогу победить '< I > shall be able to win'.

The **lexico-morphological functions** of RUMORS are limited so far to conversion between lexemes having essentially similar semantics, but differing in their part-of-speech or (for verbs only) aspectual characteristics. Thus, the aspectual or part-of-speech markers in the following three queries

разостлать:сов,

читать:III,

знать:II

cause the lexemes in these queries to be replaced, resp., by the required *perfective verb, noun,* and *adjective:*

расстилать,

чтение,

известный.

The implementation of aspectual lexico-morphological relations is based, principally, on their description in (Zaliznyak,1977). For part-of-speech relations, we have adopted, though in a very limited sense, the concept of *lexical substitutions* (Zholkovskij,Mel'chuk,1970).

If the database contains no information necessary for switching to a lexeme of the desired aspect or part of speech, RUMORS resorts to its periphrastic functions or else makes modifications in the query. E.g., the query

чиновник : I'

aimed at forming the verb corresponding to the noun *чиновник 'bureaucrat',* will be processed to produce the phrase:

делает то, что характерно для чиновника

'< He > is doing what is typical of a bureaucrat'.

2.2 Syntactical functions

RUMORS has two major types of syntactical functions: *relational* and *word-string* ones. There is also a third group of *prepositional functions* .

Relational functions may be called both in the QUERY and in the TASK mode of operation to modify the input query with regard to the *relational references* it may include. There may be references to dependency relations, where the node specified by the query acts either as dependent *(D-references)* or as governor *(G-references),* and to anaphoric relations *(F-references).* D- and G-references may contain embedded relational references, so that in the general case each reference present in the input query corresponds to a more or less complex fragment of the dependency and anaphoric structure this query is part of.

The job of the relational functions is to ensure fulfilment of the requirements for syntactical government and agreement which may be imposed on the word-form specified by the query by the dependency and anaphoric relations this query has references to. This involves extracting such requirements from the references in the query, reconciling them with each other (if there are two or more references dictating conflicting requirements), and then modifying the initial query to fit them: choosing the correct preposition or conjunction (the empty one, if needs be) to accompany the goal word-form, and altering, as required, the inflexional and part-of-speech categories within the query. E.g., the query

решение R D2(зависеть)

describing the noun *решение 'decision'* as the syntactical object of the verb *зависеть 'depend'* will produce the prepositional combination:

от решения '< depend > on < the > decision'.

Word-string functions are specific to the TASK mode of operation only. Their peculiarity is that they include some analysis-like operations making it possible to locate and process simple prepositional, adjectival and noun phrases, even if the input sequence of queries has no syntactical marking.

To be more particular, word-string processing consists in examining the queries of the input sequence one by one until the query under examination is found to answer our definition of the end of a word-string. During this examination, each query is checked for information relevant to agreement and prepositional government, and the inflexional and

part-of-speech categories pertaining to such information are integrated into a special *word-string query (w-query)*. After the end of the word-string has been located and examined, the w-query obtained is, in standard cases, made common to all of the individual queries within this word-string. Thus, the sequence of **morphologically empty** queries for lexemes

в, весь, наш, галактика 'in, all, our, galaxy'

will be processed to produce the prepositional phrase:

во всей нашей галактике

'in the whole of our galaxy'.

Some types of word-strings, e.g. those containing cardinal numerals, have to be subjected to more elaborate operations.

If a query within a word-string contains relational references, the requirements imposed by these are given priority over the requirements extracted by word-string functions, so that the latter provide a sort of default.

Prepositional functions are employed in both modes of operation, if the word-form or word-string being processed is to be preceded by a preposition. Thus, if the preposition in question denotes location, direction or source, the noun it is meant to accompany is checked for having lexical preferences in this respect. This helps to account, e.g., for such idiomatics as

на улице 'in the street'

vs.

в переулке 'in the side-street'.

Other prepositional functions serve to add the prothetic *н* to personal pronouns after prepositions imposing this requirement, to choose the contextual form of the preposition if it has more than one of them, etc.

3 Database

The database used by RUMORS has been derived, primarily, from (Zaliznyak,1977) which provides information on inflexion and aspectual conversion for about 100 000 lexemes. We used a computerized version of (Zaliznyak,1977) made available to us by the Department of the machine pool of the Russian language (Institute of the Russian language of the Russian academy of sciences, Moscow). By now, however, our database is appreciably different from its source.

Aside from various minor modifications, we have taken advantage of the fact that it is not unusual for inflexional information characterizing a lexeme to correlate with some components of this lexeme's word-structure. Such components have been organized into a dictionary of their own, the information associated with each of them included in their respective entries, and all the corrresponding lexemes removed from the database.

The database has thus been reduced to less than one fifth of (Zaliznyak,1977), still affording correct morphological processing of all of the 100 000 lexemes listed in (Zaliznyak,1977).

Moreover, so far as lexemes with standard morphological characteristics go, they can now be processed correctly, even if they are newly-coined or occasional (and do not have therefore dictionary entries of their own). As (Zaliznyak,1977) may be trusted to contain all non-standard lexemes, the inflexional and aspectual information in the resulting database very nearly covers the whole of the Russian vocabulary.

The situation is different with information to be used in part-of-speech conversion and syntactic processing, for it is not provided in (Zaliznyak,1977). This information is now also being added, but in this respect, the database is far from completed and has as yet only experimental value.

References

[1] L.S.Modina, L.S., Shalyapina, Z.M. (1994). The JaRAP experimental system of Japanese-Russian automatic translation (submitted for COLING 94).

[2] Zaliznyak, A.A. (1977). Grammaticheskii slovar russkogo yazyka (The Grammatical Dictionary of Russian). - Moscow.: Russkii yazyk, (in Russian).

[3] Zholkovskij, A.K., Mel'chuk, I.A. (1970). Sur la synthèse sémantique. *T.A.Information*, No.2.