# Russian Word Prediction with Morphological Support

**Sheri Hunnicutt[*], Lela Nozadze[†], George Chikoidze[‡]**

**Abstract**

A co-operative project between two research groups in Tbilisi and Stockholm began in 1996. Its purpose is to extend a word predictor developed by the Swedish partner to the Russian language. Since Russian is much richer in morphological forms than the 7 languages previously worked with, an additional morphological component, using an algorithm supplied by the group in Tbilisi, is seen as necessary. It will provide inflectional categories and resulting inflections for verbs, nouns and adjectives. The correct word forms can then be presented to the user of the word prediction system in a consistent manner, allowing the user to easily choose the desired inflectional word form. At present, the work with the classification of verbs is complete. The algorithm is also being used to automatically tag the large lexicon used in the word predictor with inflectional classes.

## 1. Introduction

Between 5 and 8 percent of the population has serious specific reading and writing difficulties that are often referred to as dyslexia. Teachers in schools have been aware for many years of this specific functional problem although the explanations for the source of the problem have varied widely. During recent years, most researchers have agreed on a biological cause, and dyslexia is now seen as a type of language disability.

In vocational situations, dyslexia leads to serious problems. As we depend more and more on computers to aid us in our work, the ability to read and write becomes more necessary. Even if computers are not involved, many working positions depend upon the ability of a person to handle written material, from taking notes after a telephone conversation to writing reports.

There are also others groups of persons who have difficulty in reading and writing. One group is composed of persons with motoric disabilities who write very slowly. Persons who have aphasia may also have such difficulties. Second language users also often experience reading and writing problems - these may be persons who have immigrated or persons who are deaf and have sign language as their first language.

In many cases, the introduction of the computer in the workplace has made reading and writing difficulties more obvious. But the computer also provides a possibility to aid persons with such difficulties. Training programs can help a person

[*] Dept. of Speech, Music and Hearing, KTH, Dr. Kristinas väg 31, S-100 44 Stockholm, Sweden, e-mail: sheri@speech.kth.se

[†] Guest researcher from Dept. of Language Modelling, Tbilisi, e-mail: lela@speech.kth.se

[‡] Dept. of Language Modelling, Inst. of Control Systems, Academy of Sciences, Tbilisi, Georgia, e-mail: gogi@gw.acnet.ge, gogichikoidze@yahoo.com

perfect his/her language abilities together with the help of special education. However, many persons will continue to need support, particularly in writing. It is therefore necessary to develop programs with functions to aid in writing which can help users to spell correctly, to choose among possible desired words and to correct some grammatical mistakes. A research group at KTH has worked for many years with the second of these. Programs that carry out this function are called word predictors. A word predictor suggests words while a person is writing, either from the previous word or from the first letter(s) in the current word. If needed, synthetic speech can be used to read out the words in the list of choices and to re-read the text that has just been written as an aid for continued writing. This will help users to find the word they want without having to type all of it and will also aid in spelling correctly and, in some cases, writing more quickly.

## 2. The collaborative effort

A collaborative effort between the Dept. of Speech, Music and Hearing at KTH and the Institute of Control Systems at the Academy of Science of Georgia in Tbilisi began in 1996. The purpose of the work is to extend word predictors developed at KTH [4] to the Russian language. Since Russian is much richer in morphological forms, i.e., nouns, verbs and adjectives have many more possible inflectional endings than in the languages previously developed, it was deemed important to include an additional morphological component, to be the responsibility of the group in Tbilisi.

In the early years, a working word prediction component for Russian was developed without the morphological component. The main modification to the (ideally language-independent) probabilistic predictor itself was to adapt it to function with Russian character coding systems and to provide it with a Russian language database. In order to construct this database, an extensive text corpora (2.3 million words) in the Russian language was collected. Evaluations of the algorithm gave good results, comparable to the Swedish. At the Department of Language Modelling in Tbilisi, during this time, researchers introduced special operations for constructing word forms from a word's morphological components.

More recently, a Russian grammatical text "tagger" was located in Mexico [3] and converted for use with the frequency-based word predictor, Prophet. Using this tagger, unigram and bigram lexicons (lexicons with either single words or word pairs) containing 10,000 words each were extracted from the text corpora constructed earlier and marked with their grammatical category [2]. In this context, the morphological analysis work of the group in Tbilisi has once again come into use in order to be able to mark these words with the correct inflection categories, and to subsequently be able to expand a root form to all possible forms for choice by the user of the word prediction program. In parallel, the word prediction program is being extended with a new algorithm at KTH to allow for a more complex look-up table and a 2-step presentation of morphological forms that will make it easy for the user to locate the desired form.

## 3. Morphological synthesis

The list of word forms supplied in the prediction program will be filled in automatically by the generative morphologic processor "Basic form→Paradigm" (BfP). As suggested in its name, the system responds with a complete list of corresponding paradigm members given any basic form of verb, noun or adjective. The successive steps of processing are mirrored in the flow chart below (Fig. 1).

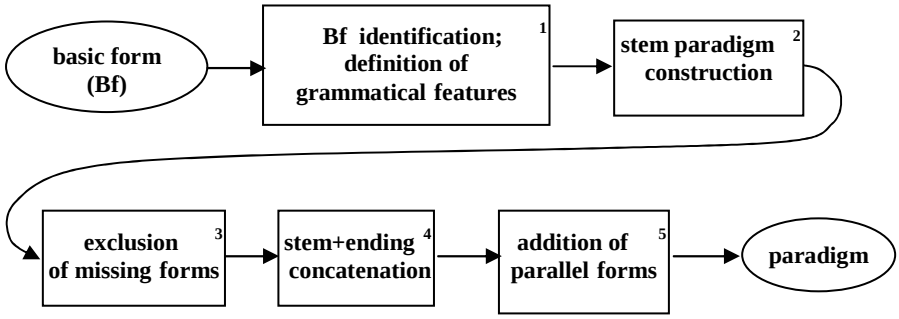



**Fig. 1 The flow chart depicts the course of paradigm generation by the BfP morphological processor.**

The first block functions as a dictionary represented as a hierarchically organized list of basic forms. The lowest level of this hierarchy comprises sublists that include those basic forms with identical paradigmatical characteristics. So, for example, all basic verb forms (infinitives) are divided according to their types according to [5]. Further, each sublist is divided according to infinitive endings (e. g. the most regular Russian verbs belonging to the 1st type may end in *-ать*, *-ять* or *–еть* e.g. *чит-ать*, *си-ять*, *пестр-еть* respectively); the latter sublists are divided once more giving rise to final sublists, or resulting categories, of basic forms, each of them having identical grammatical characteristics. So, for example, the verbs of the 5$^{th}$ type (*бежать*, *лежать*, *надлежать*, *кричать)* all have - *ать* as their infinitive ending, the same stress position scheme and identical characteristics of aspect (imperfect) and transitivity (intransitive). However *бежать* has the peculiar (for 5$^{th}$ type) ending *-ут* for third person plural, *лежать* changes the root vowel *е* to *ё* in the verbal adverb form, and *надлежать* is impersonal. Only *кричать* has no such irregularities. As a result, all four belong to different basic form categories for our purposes, the first three creating three separate categories, while the fourth is included in a category with more than 30 other basic forms corresponding to the quite regular paradigms. The resulting effect of this approach is that the system can dispense with all morphological information in the canonical dictionaries.

Using this information acquired in the first step of processing, subsequent steps are taken in correspondence with the flow chart. In Block 2, all stems of output paradigm members are constructed; in Block 3, the composition of this paradigm is

defined by excluding all superfluous members; in Block 4, the stems generated earlier are concatenated with corresponding endings; and, lastly, in Block 5, possible alternative variants are added to the paradigm members created in this way (if such exist).

Stems of Russian words undergo quite various transformations in the frames of their paradigms. For example, the verb stem may change in the forms of present tense (e.g. *стриг-у*, but *стриж-ёшь*; *похищ-у*, but *похит-ишь*). Or there may be different stems for present and past tense. These transformations of the initial infinitive stem are carried out by the second block of the scheme.

The composition of the paradigm, which is displayed in the third block of the diagram, quite frequently deviates from regularity as well. The most regular rule in this domain is the lack of passive voice participles in intransitive verb paradigms. But even this most strict regularity is violated by verbs such as *достичь/достигнуть* or *управлять*. Another frequent case is that verbs of imperfect aspect that, as a rule, must have a "verbal adverb" in the present tense, actually lack it (e.g. *бежать*).

The component of the system corresponding to the last (fifth) block of the flow chart deals with parallel forms and whole paradigms**.** Sometimes such alternatives accompany a single member of a paradigm, sometimes, some of members, but in quite many cases this parallelism spreads over a whole paradigm.

The BfP system has been completed as yet only for Russian verbs. It is implemented by means of a morphologic net representation [1] and its interpreter (in the C-programming language).

## 4. Construction of inflection table

The list of verbs from Zaliznjak, sorted according to type, are then grouped into classes with identical inflection sets. It is then possible, for each of these classes, to reduce the information to a unique set of inflections. However, this list would be too long to show on a computer screen for the user of the word prediction program. For this reason, the participles, which themselves have many inflections, are given in a second list. That is, the first step in choosing a particular participial form is the choice of one of the four participle types: present active, present passive, past active, past passive. This choice will be made by choosing the first member of the paradigm, which is common and easily recognizable. Once this choice has been made, the second step will be to choose the particular form desired. The two existing participle types for verbs of Type 15 ending in *–ть* and having the unique set of inflections #1 is shown below:

V15.1 ть ть ну нешь нет нем нете нут нь ньте л ла ло ли вший… в вши

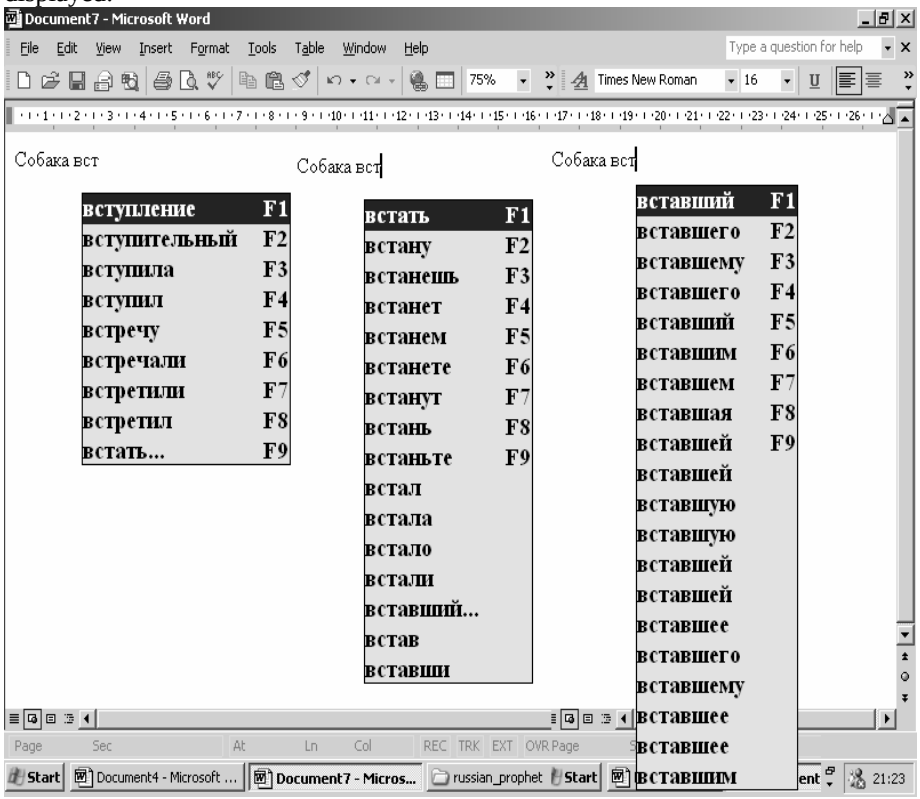
V15.1 вший ий его ему его ий им ем ая ей ей ую ую ей ей ее его ему ее ее им ем ие их им их ие ими их

A few verbs which belong to this class are встать, застать, устать, застрять, привстать, and достать.

## 5. Display of paradigm in Prophet

When a base form is chosen in the word prediction program Prophet, the program looks up the entry in the lexicon to determine which paradigm it belongs to. For example, if the user chooses the verb встать from the list of predicted words(see entry followed by three dots in the leftmost prediction window below), the lexical entry will have the following form:

Then, referring to the inflection table, the program will extract the entry shown above in the previous section, and will display it in two steps as shown below in the middle and rightmost prediction windows. These windows will appear successively, not simultaneously as shown. When the form followed by three dots in the middle prediction window is chosen, the third (rightmost) prediction window will be displayed.




The inflections in the prediction list will always have the same order, so that if the user wants вставшая which is the past tense, active voice, singular, feminine form of the participle of the verb above, this form will always be in the8th place in the list of past, active participles. The desired form is then easy for the user to locate.

## 6. Conclusions

To date, the algorithm for expanding the root form of verbs into their inflections has been completed. As verbs are the most complex word class morphologically, the work that has now been completed suggests a successful completion with the remaining inflectible words. We foresee that this system will have a wide range of applications besides word prediction, including morphological analysis/synthesis, text annotation, spell checking, dictionary look-up and language teaching.

### 7. References


[1] Chikoidze, G. (1998) "Net representation of reversible morphologic processor". *Proceedings of the Second Tbilisi International Symposium on Language, Logic and Computation*.

[2] Chikoidze, G.B., Nozadze, L.F., Javashvili, N.G. & Dokvadze, J.A. (2002) "Automatic Russian Spelling Dictionary (Verb component)," *Proceedings of the International Workshop Dialogue 2002*, Moscow (Protvino), pp. 529-539.

[3] Gelbukh, A. & G. Sidorov, G. (2002) "Morphological Analysis of Inflective Languages Through Generation," *J. Procesamiento de Lenguaje Natural*, No 29, Spain, pp. 105-112.

[4] Hunnicutt, S. & Carlberger, J. (2001) "Improving Word Prediction Using Markov Models and Heuristic Methods." *Augmentative and Alternative Communication, 17,255-264.*

[5] Zaliznjak, F.F. (1977) "Грамматический словарь русского языка." Изд. *Русский язык*, Москва.