On the buildup of an integrated database for the formal description of grammars for the hearers

Vadim Kasevich[1], Iuliia Menshikova[2], Maria Khokhlova[2], Elena Shuvalova[2], Anna Lastochkina[3]
[1]Faculty of Asian and African Studies, SpbU, Russia
[2]Faculty of Philology, SpbU, Russia
[3]Faculty of Liberal Arts and Sciences, SpbU, Russia
https://doi.org/10.36505/ExLing-2016/07/0016/000275

Abstract
Grammars for the hearers often significantly differ from those for the readers as traditional orthographic notation of wordforms is unable to fully represent the actual expression of the morphological categories and, consequently, the real composition of the paradigms. As a first step for the construction of a grammar for the hearers, one needs a database containing the information on the spoken (phonological) expression of the morphological units. At present the part of the database with the information on Russian nouns is completed. The subjects in the database are Russian noun forms of different declensions and accent paradigms expressing all the types of the stem endings that are able to shape the actual spoken realization of a form.
Key words: linguistics, Russian language, grammar, phonetics, morphemics.

Introduction
The idea of the project is based on two articles published in the 1970s: L.V. Bondarko, L.A. Verbitskaya "On Phonetic Characteristics of Post-tonic Vowels in the Modern Russian Language" and L.V. Bondarko, L.A. Verbitskaya, M.V. Gordina, L.R. Zinder, V.B. Kasevich "Styles of Pronunciation and Types of Pronouncing". The experiments on which these publications were based showed, in particular, that native speakers do not distinguish "by ear" such word forms as, for example, новая, новое: they merge into новая. And it is not a singularity, because such "merges" are found in many different segments of the system of the modern Russian language.

Baudouin de Courtenay was first to call the problem of describing the grammar of a language on the basis of oral (primary) speech one of central fundamental problems of descriptive grammar in particular and of theoretical linguistics in general. However, more than a century after the publication of Baudouin's works this problem remains unsolved. It explains the academic novelty of this project. For a long time solving this problem was considered problematic, because it required having developed and application-proven phonological and grammatical

theories. Present-day linguistics in Russia has all the prerequisites for a systematic description of the grammatical structure of the contemporary Russian language on the basis of its oral form, and the problem of creating this description is of great current interest.

The authors are not aware of any Russian or foreign research teams that would work on the problems raised in this paper. At the beginning of the XXth century there existed an international scholarly journal LE MAITRE PHONETIQUE, where all publications were printed in phonetic transcription. However, it was a purely empirical project the aim of which was to popularize the usage of transcription.

Methodology

The specific problems that are to be solved within the project are the development of two basic problematic areas. The first one is the creation of databases that would reflect changes in inflectional paradigms of Russian words that depend on their sound/orthographic codes. The second one is to reveal shifts in the system of Russian morphosyntax caused by this recoding.

Using the projected databases will allow effectively establishing basic trajectories of changes in paradigms after the change of the code (modality) of the plane of expression of linguistic units. In order to solve the formulated problem we use methods of classical structural linguistics with its focus on revealing formal paradigms that consist if oppositive word forms; categorical analysis; neutralization of oppositions in specific contexts etc. The formal paradigms that are analyzed are seen as semantisized structures, where the plane of expression and the plane of content are inseparable, and shifts in semantics normally correlate with shifts in the content plane, and vice versa. Considerable attention is given to the exploration (both theoretical and experimental) of the category of neutralization in its complex relationship with the category of homonymy.

The expected general outcome of the methods and approaches briefly described above is a model that would allow tracing all the changes of the language system that it undergoes in the transition from orthographically oriented to phonologically oriented representation.

Results and Discussion

Within the framework of the project we have created a prototype of the database filled with word forms of different parts of speech that allows tracing consistent patterns in the reduction of paradigms caused by transition from orthographic to phonological code. Working on the database will allow determining trajectories connecting “orthographic”

and "phonological" word forms and, consequently, correlate the grammar of the speaker and the grammar of the hearer.

The first stage of the project is data collection and presentation of data in the frame of the existing database. It will be build "around" separate inflected parts of speech (nouns, adjectives, numerals, pronouns and verbs). At the same time, we are going to use the results of database processing to prepare material for perceptive experiments.

The results of the project are to be on open access, so choosing the data format was an important decision. We have selected the XML format as the most universal and well adapted to future conversion for the developing database. Below is an example of a fragment of XML representation of the lexical item «окно».

```
<entry id="n50" author="yum" time="2016-05-28">
<word>окно</word>
<orth>окно</orth>
<grammar>1d*</grammar>
<accent>B</accent>
<url>http://ru.wiktionary.org/wiki/окно</url>
</entry>
```

We have chosen the platform Microsoft SQL Server for database maintenance because of its reliability, scalability and productivity.

The chosen database format is based on client-server architecture; the server side provides most functionality while the client presents a graphic interface for the users. The client applications contact the server via the standard HTTP protocol. The server part is build up from small parts called servlets that allow for the composition of all servers from modules. Each servlet provides functionality, e.g. the database access, search, morphological analysis and connection to various corpora if required. Thanks to different commands it is possible to receive various results corresponding to queries (including combined queries). For example:

- a paradigm member or the initial form in orthography;
- a paradigm member or the initial form in phonological transcription;
- a grammatical characteristic on one morphological category;
- a grammatical characteristic on a given set of morphological categories;
- information on the homonymy of inflectional elements in orthography;
- information on the homonymy of inflectional elements in phonological transcription;
- information on the allomorphism of inflectional elements in orthography;
- information on the homonymy of inflectional elements in phonological transcription etc.

In the database there is search with wildcards support (of the language of regular expressions), so it is possible to search for parts of words or expressions. At present we are working on creating algorithms of data processing for the database of the selected type on the basis of a completely filled fragment of the nouns database. The objects of the database are the word forms that represent Russian nouns of different types of declensions and accent paradigms and demonstrate all the types of stem endings that can influence the phonetic image of the word form. The fields of the database contain information about the orthographic and phonetic image of a word form, about all of its morphological characteristics, variability of morphological forms and accent patterns, inflection indexes and accent types. Different fields contain the orthographic and phonetic images of stems and inflectional affixes included in each word form.

References

Bondarko L.V., Verbitskaya L.A. 1973. On Phonetic Characteristics of Post-Tonic Inflexions in the Contemporary Russian Language. In Problems of Linguistics, No 1, 37-49.

Bondarko L.V., Verbitskaya L.A., Gordina M.V., Zinder L.R., Kasevich V.B. 1974. Styles of Pronunciation and Types of Pronouncing. In Problems of Linguistics, No 2, 64-70.