Rajna Dragićević, Yury Makarov, Daria Ryzhova,
Yulia Shapich, and Ekaterina Yakushkina

A NEW SERBIAN-RUSSIAN DICTIONARY

Abstract The paper presents a New Serbian-Russian dictionary and the main principles of its development. We use the most recent explanatory dictionary of Serbian, published by the Serbian Academy of Sciences and Arts in 2018, as a starting point. However, we refine both the word list and the entry structure to meet the requirements of a bilingual edition. We consult text corpora of modern Serbian to check frequencies and define contexts of word usage. The paper describes the criteria for word list formation, discusses the problem of lexical equivalence in closely related languages, and articulates our guidelines for representing polysemy. The dictionary will be published in both traditional print format and as a website. The digital version will be implemented using the OnLex platform; we detail its functionality and highlight the new possibilities it offers to editors and potential users.

Keywords Serbian; Russian; dictionary; online lexicography; OnLex

1. Introduction

The number of various electronic resources, such as corpora, dictionaries, and electronic databases for various languages, including Slavic, is continuously and rapidly growing. Despite this, there is currently no sufficiently complete and up-to-date dictionary which would be capable of fulfilling the needs of modern language users interested in Serbian and Russian. The most recent bilingual dictionary was published in 1957 by Iliya Tolstoy (expanded edition — 1970) (Tolstoy, 1970). Since then, not only have the vocabularies of both languages changed but also the approaches to writing and publishing dictionaries as well as lexical analysis in general.

To fill in this gap, we are creating a New Serbian-Russian dictionary. It is an ongoing project, implemented at the Department of Slavic Philology of Lomonosov Moscow State University in collaboration with the Department of Serbian Language at the University of Belgrade. For now, entries for words starting with the letters A, B, V, G, and D are already complete (approximately 10,000 entries in total), and those starting with the letters Đ, E, Ž, and Z are in progress. The estimated size of the complete dictionary is 80,000 entries.

The New Serbian-Russian dictionary is being created based on the most recent edition of the explanatory dictionary of the Serbian Academy of Sciences and Arts (RSJ, 2018). In compiling the dictionary and clarifying the meanings of words, text corpora, such as SrWac (2014) or PDRS (2022) are actively used. The dictionary is being developed in parallel in the traditional text format and in an electronic version on the OnLex platform (Makarov, 2024).

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

The article is organized as follows. Section 2 will present the principles of dictionary formation. Section 3 will be dedicated to the problem of lexical equivalence and the form of representing polysemy. Section 4 will introduce the digital version of the dictionary. Finally, Section 5 discusses the prospects for the development of the project.

2. Forming the Word List

We use the vocabulary, overall entry structure, and the meaning representation of the most recent Matica Srpska explanatory dictionary of Serbian (RSJ, 2018), comprising about 80,000 lexemes, as a starting point for our project. However, every part of a dictionary entry taken from RSJ (2018) is critically evaluated and the word list is recompiled taking into account the audience for which the new bilingual dictionary is intended.

The dictionary is devised primarily for Russian-speaking users. Its main purposes are to help the users read classic Serbian literature and contemporary texts of various genres, as well as to provide Russian learners of Serbian with the information necessary for the appropriate learning and everyday use of Serbian lexemes. These goals determine the principles of dictionary compilation, which we formulate as follows:

1. Lexemes that, according to the lexicographer's assessment, belong to the common lexical stock, are used in different functional styles, and are sufficiently frequent, are included in the Serbian-Russian dictionary as a matter of course.

2. Lexemes that are found in culturally significant literary works, both folk and authored, written in Serbian, are also included in the bilingual dictionary, even if their frequency in modern language is not high, e.g., *gradonosac* 'a cloud that carries hail'.

3. Words that belong to the terminological apparatus of a particular subject area are included in the dictionary if they are known to the modern educated speaker of the language; for more details on the principles of including terminological vocabulary, see Korolkova (2023).

4. International names of plants and animals are added to the dictionary if they meet two conditions:

    a) the word has a different form in Russian, e.g., Serbian *ginseng* 'ginseng (Latin *Panax ginseng*)' corresponds to Russian *žen'šen'*, which looks and sounds differently.

    b) a plant or an animal denoted by the term is important to either Serbian or Russian culture and everyday life, i.e., the term frequently appears in Serbian or Russian discourse.[1]

[1] The same criterion applies to frequently used proper names which sound different in Serbian and Russian, cf. Serbian *Beč* 'Vienna,' corresponding to Russian *Vena.*

5. When selecting certain words, a word-formation criterion is used: a word is included in the dictionary if it serves as a productive base for various kinds of derivatives, even if the word itself is very rarely used in modern language (compare *Gestapo* 'Gestapo' — *gestapovski* 'harshly', *gonetati* 'to guess' — *zagonetka* 'riddle').
6. When including proper names in the dictionary, their semantic potential is also considered: if a proper name forms new meanings and can be used as a common noun, this is considered sufficient reason for including it in the dictionary (e.g., *Golgota* 'Calvary' and *golgota* 'anguish, agony').
7. In addition, the RSJ dictionary is supplemented with new words, primarily recent borrowings from English, e.g., *gejmer* 'gamer' or *globalizam* 'globalism'. Dictionaries of neologisms, foreign words, and Anglicisms, such as Prćić et al. (2021) or Otašević (2008), are used as sources of such vocabulary, and the extent to which they are adopted, which is important for making the decision (not) to include a lexeme in the dictionary, is determined by their frequency.

The frequency of a word in modern Serbian is assessed using corpora of texts written in modern Serbian, such as PDRS (2022), SrWac (2014), SrpKor (2013) and some others. These corpora represent collections of modern internet texts collected from *.rs* domain sites over various periods between 2002 and 2022. All the corpora include a lot of data: the volume of the largest of the corpora, PDRS, is about 715 million tokens. We believe that such collections of texts provide a reliable representation of the degree of a word's usage in the modern language; however, they certainly do not allow us to define the significance of the word for the general cultural and, especially, literary fund, as they do not take into account works written before the beginning of the 21st century. This lack of objective data is compensated for by the linguistic competence of the dictionary compiler since the word list is formed by a lexicographer who is a native speaker of Serbian.

3. Lexical Equivalence and the Problem of Representing Polysemy

When working with closely related languages, such as Serbian and Russian, the problem of lexical equivalence becomes particularly acute. Many words have a common origin and a very similar set of meanings. For example, the Serbian noun *bogatstvo* corresponds to the Russian cognate *bogatstvo*, and both words denote owning much property, as well as spiritual wealth (cf. Serbian *duhovno bogatstvo* vs. Russian *duxovnoe bogatstvo* 'spiritual wealth') and great variety (cf. Serbian *bogatstvo boja* vs. Russian *bogatstvo krasok* 'a riot of color'). However, in most cases, careful analysis of collocations allows us to identify discrepancies in the use of tentative equivalents. For instance, the Serbian collocation *čitavo bogatstvo* 'absolute fortune' corresponds to the Russian expression *celoe sostojanie*, and not $^{?}$*celoe bogatstvo*, as one might expect.

Large text corpora, including parallel ones, such as Serbian-Russian and Russian-Serbian fragments of the RNC (2024) or InterCorp (2023), provide us with the opportunity to track subtle differences in superficially similar collocations and illustrate them with the most indicative, i.e., the most frequent and stable (idiomatic) examples. In addition, the possibility of publishing an electronic, not just a paper version of the dictionary allows us not to save space at the expense of illustrative material. New opportunities also change the general principles of developing a dictionary entry for a bilingual dictionary: a large amount of illustrative material allows for a focus on more precise explication of meaning, rather than on the selection of more precise translation equivalents, cf. Gudkov (1974).

In developing the New Serbian–Russian Dictionary, we are aiming for the comprehensive description of the semantics of the Serbian word. This corresponds to our general orientation towards Russian-speaking users, for whom Serbian is not a native language: it is precisely the representation of polysemy and possible contexts of word usage that facilitates subsequent correct use of Serbian lexemes and their adequate translation.

If a Serbian polysemous word has a Russian translation corresponding to it in the majority of senses (e.g., Serbian *gluv*, Russian *gluxoj* 'deaf'), we do not just mention it in the entry. Instead, we try to present the entire structure of polysemy, i.e., to list all meanings of the Serbian word, illustrate each meaning with typical examples, and also include the synonyms of the translation specifically in this kind of usage. For instance, (1) illustrates the adjective *gluv*, which has several senses. In each sense, it modifies different nouns and has distinct translations. This way of representing polysemy allows us, among other things, to highlight the difference between the Serbian word and its primary translational equivalent (in this case, the cognate Russian *gluxoj* 'deaf,' which is given as a translation for all the meanings except 2d).

(1) *A part of the entry for the Serbian* gluv *(the orthography of the dictionary is preserved; Serbian expressions are given in bold):*

1a. глухой, слабослышащий 'deaf, hard of hearing'; ~ **на десно уво/ухо** глухой на правое ухо 'deaf in the right ear'

1b. fig. глухой к чему-то, равнодушный 'deaf, immune'; ~ **према нечијим жалбама, захтевима** глухой к чьим-то жалобам, требованиям 'deaf to someone's complaints or demands'

2. fig. a. глухой, тихий, без звуков; полный (о тишине) 'quiet, dead, total (silence)'; **~а ноћ** глухая ночь 'dead of night', **~а тишина** глухая тишина 'hollow silence'

2b. глухой, далекий, заброшенный, захолустный 'remote'; **~о село** глухая деревня 'remote village'

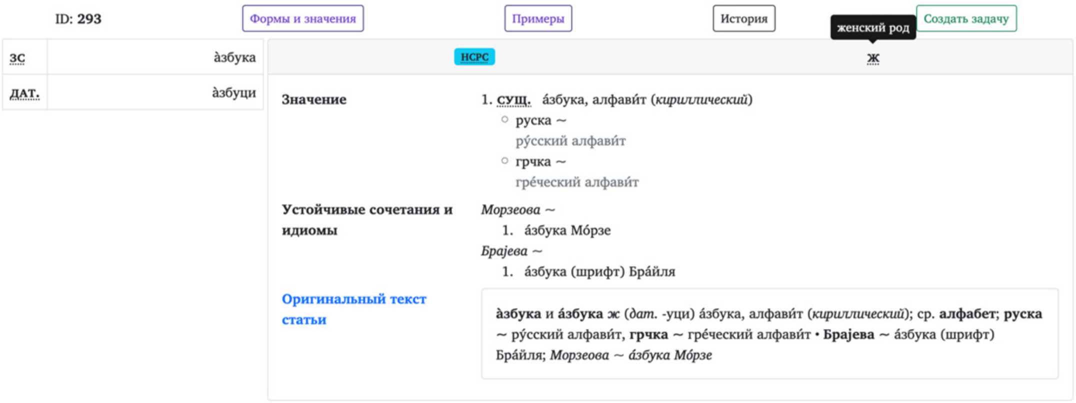
2c. глухой, тихий, неоживленный 'deserted, quiet'; **~а улица** глухая улица 'quiet street'

> 2d. тихий, нешумный, без песен, веселья (о празднике) ‘quiet, noiseless, small’; ~ **свадба** тихая свадьба ‘quiet wedding’
>
> 2e. глухой, тихий, приглушенный, неясный ‘quiet, muffled, dull, hollow’; ~ **шум** глухой шум ‘dull sound, muffled noise’, ~ **глас** глухой голос ‘hollow voice’, ~**и кораци** глухие шаги ‘hollow footsteps’

The issue of explicating meaning (primarily of polysemous words but not only them) in creating a bilingual dictionary for closely related languages is the most crucial, and in our view, it has not yet received a definitive solution. As new data accumulate, we refine and revise our principles for organizing dictionary entries and representing polysemy, and, just as importantly, we notice various patterns in the structure of the lexical systems of closely related languages that were not visible before.

4. OnLex Platform

Aside from publishing a hardcover edition of the dictionary, we are also developing its online version, powered by the OnLex platform (Makarov, 2024). The main motivations include being able to make constant amendments after the physical copy is ready and increasing the availability of the dictionary. Not only the OnLex platform is a publishing solution, allowing one to search within different parts of the entry and therefore making the dictionary useful for Serbian learners of Russian as well, but it is also advanced dictionary writing software, providing a lexicographer with a means of editing every component of the entry via an intuitive user interface. Finally, the online dictionary can render supplementary media files (e.g., we are planning to add audio files illustrating the pronunciation of lexeme paradigms since Serbian has variable lexical stress) and include more examples.



ID: 293

зс	àзбука
дат.	àзбуци

Значение
1. сущ. áзбука, алфави́т (*кириллический*)
- руска ~
  ру́сский алфави́т
- грчка ~
  гре́ческий алфави́т

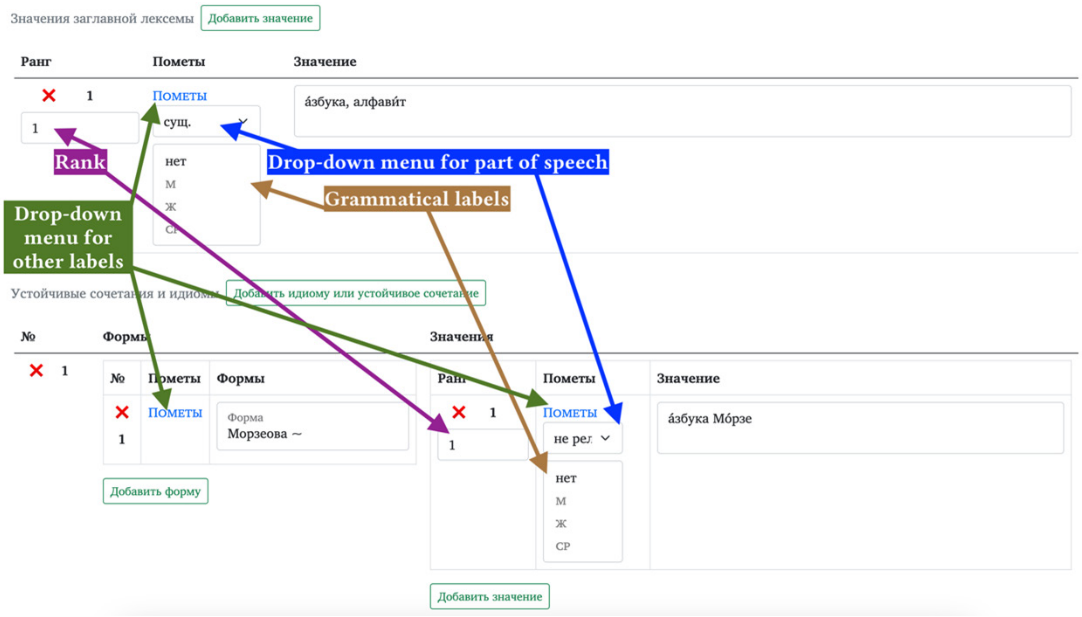
Устойчивые сочетания и идиомы
Морзеова ~
1. áзбука Мо́рзе

Брajeва ~
1. áзбука (шрифт) Бра́йля

Оригинальный текст статьи
àзбука и **áзбука** ж (*дат.* -уци) áзбука, алфави́т (*кириллический*); ср. **алфабет**; **руска** ~ ру́сский алфави́т, **грчка** ~ гре́ческий алфави́т • **Брajeва** ~ áзбука (шрифт) Бра́йля; *Морзеова* ~ *áзбука Мо́рзе*

Fig. 1: Screenshot of the entry *àzbuka* ‘alphabet’ as rendered by the OnLex platform. In the upper left corner word forms are given (the main form, labelled ‘зс,’ and a dative form, labelled ‘дат.’); note that in the original text of the entry (‘Оригинальный текст статьи’), i.e., in the print edition, the dative form is shortened. To the right of the word forms, there are senses of the lexeme (‘Значение’), illustrated with two examples and their translations, and idioms / fixed expressions (‘Устойчивые сочетания и идиомы’) along with their meanings. Every abbreviation, be it a form type (зс = headword, дат. = dative), grammatical information (ж = feminine, сущ. = noun) or a usage label, is given a pop-up tooltip, i.e., the full form appears on the screen if the user positions a cursor over the abbreviation.

As shown in Figure 1, every part of the entry is represented separately, i.e., the OnLex platform treats it as a distinct entity and stores it in a separate cell in the database. Such an approach allows us not only to render each entry component with a distinct set of typographical conventions but also to automatically generate a layout required for the paper edition. On top of that, the database structure treating every part of the entry as a separate entity (rather than representing the whole entry as a string of characters with no markup) allows for flexible search queries. For example, it is possible to search not only within headwords but also within idioms, examples, word forms (e.g., dative forms), etc.



Значения заглавной лексемы

Ранг	Пометы	Значение
1	сущ.	а́збука, алфави́т

Устойчивые сочетания и идиомы

№	Пометы	Формы
1		Форма Морзеова ~

Ранг	Пометы	Значение
1	не рег.	а́збука Мо́рзе

Добавить значение
Пометы
нет
м
ж
ср
Добавить идиому или устойчивое сочетание
Добавить форму
Добавить значение
Rank
Drop-down menu for other labels
Drop-down menu for part of speech
Grammatical labels

Fig. 2: Screenshot of the entry editing interface used by the OnLex platform with the main elements labelled. Ranks are used to order entry elements. Each label is chosen from predefined lists, which minimizes the chance of misprints.

The OnLex entry editing interface is partly shown in Figure 2. Note that no programming or technical skills are required to fill in the fields of this interface – editors either have to type unformatted text or select the required options from drop-down menus with predefined labels. All the changes are implemented and available for users right after the editor clicks on the save button and the system keeps the record of the versions of the entry. This interface is also used for writing new entries, which can be added to the online version immediately or just saved for internal use.

Among other features of the OnLex platform, there is a task tracker tailored specifically to the needs of the lexicographers working on the dictionary. Tasks can be created for specific entries and/or assigned to specific editors. These options make it possible for a logged-on editor to see that a particular entry has been mentioned in a task and when the task is assigned to a particular editor, this editor is notified. For each entry, there is a discussion board and a timeline showing who and when worked on the task.

5. Conclusion and Future Prospects

In this article, we have introduced the New Serbian-Russian Dictionary, the creation of which is associated with solving a number of practical and theoretical problems. The main theoretical problem facing the compilers of such a dictionary is the issue of the presence or absence of equivalence between translation units. This question is related to the means of representing polysemy: is it sufficient to bring the closest translation equivalent if it has a similar structure of polysemy, or should the set of meanings of a polysemous word be presented in detail in any case, supplying it with numerous illustrations? In the current version of the dictionary, the second option is implemented, but as we accumulate material, we constantly modify and refine the principles we have already developed.

Practical tasks are associated with the choice of methods for analyzing and presenting material. In analyzing the meanings of words and searching for the most illustrative examples, we use a whole range of available corpora; however, each of them reflects only the contemporary state of the language and does not include classical literature. Using the online platform OnLex contributes to more flexible work on the dictionary. It facilitates the task of adding new data and the process of changing the existing dictionary entries, as well as makes it possible to attach an unlimited number of illustrations and full paradigms. It is worth pointing out that each entry can be supplemented with audio recordings of each form constituting the lexeme's paradigm, which is important for languages with variable (i.e., not fixed) lexical stress.

Use of the methods of online lexicography opens up great prospects for further work: it is possible to reflect ongoing lexical changes in both languages via regular and systematic updates of the bilingual dictionary, to add translations of the same dictionary entries into other languages (for example, other Slavic languages), to conduct various kinds of quantitative research, and produce different visualizations (e.g., graphs illustrating relations between translation equivalents).

References

Gudkov, V. P. (1974). Dvujazychnye slovari russko-serbskokhorvatskie, serbskokhorvatsko-russkie i perspektiva ikh sovershenstvovanija [Russian-Serbo-Croatian, Serbo-Croatian-Russian bilingual dictionaries and the prospect of their improvement]. *Sovetskoe slavjanovedenie, 6*, 65–73.

InterCorp 2023: *The parallel corpus InterCorp. Release 16.* https://intercorp.korpus.cz/?lang=en. (last access: 1 June 2024)

Korolkova, M. D. (2023). Special'naja leksika v dvujazychnom slovare [Special vocabulary in a bilingual dictionary]. *Slavistika, 27*(2), 190–201.

Makarov, Y. (2024). Principles and methods of digital lexicography. *Izvestiia Rossiiskoi akademii nauk. Seriia literatury i iazyka, 83*(4), in print.

Otašević, Đ. (2008). *Rečnik novih reči* [Dictionary of new words]. Alma.

PDRS 2022: *Serbian Web Corpus PDRS 1.0.* https://javnidiskurs.rs/veb-korpus-serbian-web-pdrs/. (last access: 1 June 2024)

Prćić, T., Dražić, J., Milić, M. et al. (2021). *Srpski rečnik novijih anglicizama* [Serbian Dictionary of new anglicisms]. Filozofski fakultet.

RNC 2024: *Russian National Corpus.* Retrieved June 1, 2024, from https://ruscorpora.ru/

RSJ 2018: *Rečnik srpkoga jezika.* Matica Srpska.

SrWac 2014: *Serbian Web Corpus v1.2.* Retrieved June 1, 2024, from https://ruscorpora.ru/

SrpKor 2013: *Korpus savremenog srpskog jezika.* Retrieved June 1, 2024, from https://ruscorpora.ru/

Tolstoy, I. I. (1970). *Serbskokhorvatsko-russkii slovar'* [Serbo-Croat–Russian dictionary]. Sovetskaja entsiklopedija Publ.

Acknowledgements

We deeply thank our anonymous reviewers for their valuable questions and useful comments. The study is funded by the Russian Science Foundation, project No. 23-28-00126, https://rscf.ru/project/23-28-00126/.

Contact information

Rajna Dragićević
University of Belgrade, Belgrade, Serbia
rajna.dragicevic@fil.bg.ac.rs

Yury Makarov
University of Cambridge, Cambridge; Vinogradov Russian Language Institute, RAS; Institute of Linguistics, RAS
im562@cam.ac.uk

Daria Ryzhova
HSE University, Moscow, Russia
daria.ryzhova@mail.ru

Yulia Shapich
Lomonosov Moscow State University, Moscow, Russia
julija.sapic@gmail.com

Ekaterina Yakushkina
Lomonosov Moscow State University, Moscow, Russia
jkatia@yandex.ru