The *Russian Academic Neography* Information Retrieval Resource

Marina N. Priemysheva, Yulia S. Ridetskaya, Kira I. Kovalenko

Institute for Linguistic Studies, Russian Academy of Sciences,
199053, 9 Tuchkov pereulok, St. Petersburg, Russia
E-mail: mn.priemysheva@yandex.ru, vjs_neolex@mail.ru, kira.kovalenko@gmail.com

Abstract

Creation of electronic dictionaries and retrodigitalization are very popular trends in modern lexicography. The idea to use computer techniques in Russian neology appeared in 2013, but only recently has the *Russian Academic Neography* information retrieval resource been created. It represents both published dictionaries (annual, decadal and thirty-year dictionaries), which include about 116,000 words and collocations that had not been registered by normative explanatory dictionaries of the Russian language, and new materials that were not included in published volumes or that are being prepared for publication. Simple and advanced types of search give an opportunity to find words by various parameters (word, word component, year or time period, labels, etc.). It is also intended to include chronological and frequency parameters in the future. The aim of the *Russian Academic Neography* information retrieval resource is to represent the newest Russian vocabulary and to make it available for a wide spectrum of users.

Keywords: new-word dictionary; neography; electronic dictionary; *Russian Academic Neography*; information retrieval resource

1. Introduction

New scientific directions, such as corpus linguistics and computer lexicography, have allowed authors and publishers of dictionaries to go beyond traditional paper lexicography and discover new possibilities for creating and using vocabulary information. During the last two decades, lexicographers have searched for optimal forms and means of achieving the most convenient and productive ways of representing vocabulary, as well as going beyond the existing formats through the creation of new, interactive resources.

At the present stage, there are several trends in the presentation and use of dictionaries in electronic form. Nowadays, lexicography works can be roughly divided into electronic dictionaries (newly created computer dictionaries and online dictionaries) and retrodigitalized dictionaries (all forms of paper dictionaries converted into electronic format).

Electronic dictionaries are a very successful genre of modern computer lexicography, and appear in various forms and solutions. As a Russian language resource, they are

incredibly popular. Along with small projects, such as the online dictionary of jargon and slang (http://www.slovonovo.ru) and the popular dictionary of the Russian language (http://slovoborg.su), large electronic thesauruses catering for a variety of functional purposes and based on databases of various sizes are available: the dictionary of collocations based on the National Corpus of the Russian language (http://www.ruscorpora.ru/obgrams.html), "Database of pragmatically marked vocabulary" (http://spml.ipmip.nspu.ru/?action=main), *CrossLexica* (https://www.xl.gelbukh.com), open electronic thesauruses of the Russian language (https://russianword.net; http://ruslex-encode.ru) and many others. Of course, this trend in the development of modern computer lexicography still requires a long period of adaptation and crystallization of forms and tasks, and each project needs to find its place in the scientific paradigm: the creation and design of such dictionaries often resembles a lexicographic game, rather than a serious scientific project.

Retrodigitalized dictionaries are represented quite significantly on the Internet, but there is still a question of technical implementation. These dictionaries have various formats: from a database of a single edition (for example, https://www.slovardalja.net, https://ushakovdiction.ru, http://orfo.ruslang.ru) to databases of dictionaries of the same type (http://etymolog.ruslang.ru/) or a number of typologically diverse dictionaries (https://www.slovari.ru, http://gramma.ru/SPR/?id=1.0, http://gramota.ru/slovari/, etc.). Currently there is a tendency to create compilations of dictionaries, such as, for example, the "Historical Dictionary of the Russian Language" (http://dic.feb-web.ru/rusdict/index.htm) and "Academic Corpus of the Russian Language Vocabulary" (Lesnikov, 2019a, 2019b). The architecture, structure and interfaces of these databases are very diverse: each of the electronic lexicographic projects in Russia currently functions autonomously, and there is still an ongoing search for an optimal electronic lexicographic form.

2. General characteristics of the texts in the *Russian Academic Neography* portal

When considering the tradition of lexicographic representation of dictionaries and dictionary resources in Russia, the *Russian Academic Neography* information-retrieval resource occupies an intermediate place and this, among other things, is its originality:

- it is both a professional resource for specialists of lexicology and lexicography, and a reference resource, designed for a wide spectrum of users;
- it is a joint database of dictionaries of the same type (annual, decadal and a thirty-year dictionaries);
- it is (in the near future) an online dictionary of new vocabulary.

At the same time, the textual database of the resource is quite specific, which is determined by the traditions of Russian neography. Russian academic neography as a separate theoretical and practical lexicographic trend has existed since the 1960s. Its theoretical basis is formed in numerous works of N. Z. Kotelova, E. A. Levashov and T. N. Butseva. As was determined by Kotelova, Russian neography is represented by three types of neologism dictionaries, work on which was conducted, and continues to be conducted, by the team of the New Words Dictionaries group of the Institute for Linguistic Studies of the Russian Academy of Sciences (Leningrad / St. Petersburg).

1. *New in Russian Vocabulary* annual dictionaries, recording all the new words of a given year, including innovations of particular authors and occasionalisms (18 issues were published: 1977–1994; work on the annual dictionaries of 2010–2019 has been resumed recently). This is a series of reference dictionaries that include absolutely all the innovations of Russian speech in the focal period. "The *New in Russian Vocabulary. Dictionary Materials...* annual dictionaries are an attempt to show the flow of spontaneous language life, to demonstrate the facts of birth, change, or entry into the language of words in all their diversity. They present everything new that occurred during daily examination in the texts of ten sources (constant from year to year) in four checked months (of a given year), including the words of short-term existence and one-time use. Each annual dictionary includes about 4,000 vocabulary units" (Kotelova, 2015: 367).

2. *New Words and Meanings* decadal dictionaries record only those lexical units that entered the Russian language in a given decade and were included in the language use. *New Words and Meanings* are explanatory dictionaries, which complement large explanatory dictionaries of the literary language (such as the *Big Academic Dictionary of the Russian Language*), as "decadal dictionaries show only facts that have become the property of the language, at least for a certain time" (Kotelova, 2015: 367). Decadal dictionaries of the 1960s, 70s, 80s and 90s have been published; the last one is a three-volume book (about 1,000 pages per volume), in which the linguistic elements of Russian life of the 1990s are clearly and visually represented.

3. *Dictionary of New Words* is a thirty-years dictionary and records only the words that entered into common usage and could be included in the dictionaries of the Russian literary language. *The Dictionary of New Words of the Russian Language of 1950–1980s* is the normative explanatory dictionary of neologisms of the post-war era, which was intended to complement the explanatory dictionaries of the Russian language.

Currently, the resource of all published new-word dictionaries is about 116,000 words and collocations that were used in the Russian language in 1960–2000, but which had not been registered by any of the explanatory dictionaries of the Russian language of the 19th and 20th centuries (in comparison, the *Dictionary of the Modern Russian Literary Language* in 17 volumes includes about 120,000 words). That means that these dictionaries significantly complement all available vocabulary resources of the Russian

language, containing as many words as had been registered by the lexicographic works before.

On the one hand, each of the dictionaries of the series has its own special scientific function but, on the other hand, it also has a complementary relationship, from a historical perspective, with another type of new-word dictionary. N.Z. Kotelova noticed that "Depending on the lexicographic situation, society needs one or another dictionary of neologisms. The need, for example, to create a normative dictionary of neologisms of a significant period can be considered to be less pressing in a situation of rapidly reprinted and updated general explanatory dictionaries. Dictionaries of new words are designed to facilitate knowledge of the language, giving a description of the innovations from the various points of view: they show their internal form (first of all, the producing word), supply stylistic labels, give forms of inflection, illustrate with good examples of usage, and help with mastering the best variant among competing options. This information is also needed for translators and authors of bilingual dictionaries" (Kotelova, 2015: 370).

Also at the disposal of the new-word dictionary compilers is a fourth resource, which is not available to a wide audience: it is a bank of Russian neologisms, "including three indices: 1) words, 2) word meanings, 3) collocations. It gives an opportunity to review the entire array of neologisms, see the development of pre-existing derivational, thematic nests and series of words, the formation and degree of filling of new ones, evaluate quantitatively innovations for a given attribute (derivational, partial, structural and phraseological, etc.), compare with innovations in other languages — in general or by ranks, to see the variation or synonyms, to observe projections into extra-linguistic spheres. It fixes a point of reference for future work in the field of neology — it provides the possibility of automatic processing of neological material, the implementation of formal transformations (for example, the compilation of a reverse vocabulary of neologisms), etc." (Kotelova, 2015: 370). In other words, the bank of Russian neologisms helps to find a new language unit and define its place in the language lexical system.

Reflecting the synchronous level of the Russian language, annual dictionaries form the basis of decadal dictionaries, and each of the types becomes the historical dictionary of the Russian language of the period being described. However, the main value of a series of new-word dictionaries lies not only in the combination of historical and synchronous approaches in the lexicographic description, but also in fairly accurate dating of one or another occurrence: it is the combination of these principles that makes up the peculiarity of Russian neography.

At the present stage of the collection, recording and description of new words, the work of lexicographers has become even more complex.

Before entering the dictionary, words and word meanings must pass a multistage selection process. First, the material from the source list for the primary search is

analysed. For 2018 the list included *Komsomolskaya Pravda*, *Kommersant*, Gazeta.ru, *Rossiyskaya Gazeta*, *Vedomosti*, Lenta.ru, *Izvestia*, Rbc.ru, *RBK* (magazine), *Metro* (newspaper), and *Novy Peterburg* (newspaper). The source list is created on the basis on IndEx — an indicator calculated by the *Integrum* information and analytical system, which assesses the resource rank in the media space. "The calculation takes into account the number of publications in the media, the visibility of mentioning the object in the media, the role of the object in the publication, emotional colour of the publication and the significance of the (cited) source... The higher the indicator, the more visible the analysed object is in the media space" (https://www.integrum.ru/ratings/smi/media/jul18).

The survey of sources also includes monitoring of social networks, news feeds, popular blogs and non-professional Internet dictionaries. The initially selected lexical material is rigorously tested for novelty using the internal databases of the Institute for Linguistic Studies, as well as authoritative normative, explanatory and special dictionaries, Russian National Corpus, corpus of the Russian texts in Google.books.com, and the *Integrum* corpus of texts — the largest in the Russian Archive of texts of Russian language media. Contextual queries in the *Integrum* information-analytical system help in selecting new vocabulary that is synonymous, antonymous, hyponymic, etc. for previously found neologisms.

Modern methodological principles for the selection of lexical units for academic dictionaries of neologisms and the formation of a new-word database were developed in the early 2010s. Thanks to corpus data, it became possible to clarify the first written record of a word in Russian language texts, that is, to find out the approximate time that a word appeared in the language.

Thus, at present, the following vocabulary is available for study (classified according to time and quality parameters):

- neologisms of 1990–1999;
- neologisms of 2000–2009;
- neologisms of the last decade;
- vocabulary dated to the period of the 1990s and missed in explanatory, orthographic, terminological and other authoritative dictionaries;
- occasionalisms, individual authorial innovations, i.e. neologisms, the written record of which is unique.

The new words, new meanings, compounds and collocations that are found enter a local neological database accompanied by technical and information marks. The neologisms of the last decade are distributed by year (2010–2019) in order to create the primary word lists of the *New in Russian Vocabulary. Lexical materials.*

3. The history of the creation of the *Russian Academic Neography* information retrieval resource

New computer technologies have made it possible not only to expand the sources of new-word dictionaries, but to present the vocabulary data of academic neography in open access. Materials of all published dictionaries are represented on the website of the Russian Academy of Sciences http://iling.spb.ru/dictionaries.html.ru. The materials of the four decadal dictionaries can be found in Wiktionary https://ru.wiktionary.org/wiki/. Information about the neologisms of the last decades is being published on the web-page of the Academic Neography in the social network https://www.instagram.com/neographia.spb.

However, the needs of modern science have long dictated the transition to a new paradigm for the creation and use of dictionaries of new words: going out beyond the existing series of dictionaries makes it possible to create a resource of all neological publications and also to continue the work in the new online format. The future implementation of this lexicographic information retrieval resource will not adopt the existing principle of transition from paper format to electronic, but instead that of online format to paper, in which the paper format can be optional and diverse. This will allow us to speed up the introduction of new words into scientific circulation by representing them in the resource soon after their appearance in speech.

The idea of such a resource — the Neology Service of the Russian Language (neologia.ru) — came from the team leader T. N. Butseva (Butseva, 2013). The resource was technically developed at a very high level on the basis of a specially developed program with an original interactive interface (Dmitriev, 2013). However, the main obstacle in its creation and work was the incredible difficulty of marking up and converting 116,000 dictionary entries into the electronic database. The tasks set by the authors of the project, which were very important for Russian science, required enormous technical and human resources and have not been implemented.

Recently, the *Russian Academic Neography* electronic information retrieval resource (https://neographia.iling.spb.ru) has been developed, which continues and develops the concept of the previous resource. At present, it has reached the advanced stage of technical finalization and functions in its test mode (the main part of the vocabulary from 1977-1990s is going to be available by September 2019, in October and November it will be filled by new units for 2016-2017, and at the beginning of 2020 new materials for 2013-2015, 2018 and 2019 will be included).

4. Resource interface and functionality

The *Russian Academic Neography* resource includes a database of the published dictionaries and some unpublished materials, together with a query system. It is both a lexicographic resource and an information portal of Russian neology and neography as a whole.

The new-word database includes both the previously published lexicographical works and the new editions of annual dictionaries created by the team members. The database will be supplemented by new lexical units that were not included for one reason or another in published volumes or materials that are being prepared for publication.

In the final version the resource will include following subdivisions:

1. Information about Academic Neography.

2. Information about dictionaries and dictionary corpora.

3. Links to interesting neologisms of the current year (as a news feed); neologisms from dictionaries of previous years (period 1960–2010s); rare neologisms not represented in the dictionaries of the period before the 1960s (section "From the history of words"), as well as lexicographic and linguistic sketches and articles.

The technical implementation of the *Russian Academic Neography* information retrieval resource has been created by A. Andreev. The dictionaries are processed using a specially written program, which is based on the SWI-Prolog 7.6 development environment. Internally, the set of word entries is stored as a semantic network, with nodes corresponding to different fields in an entry, which had been identified by their formatting (font, size, etc.). TEI encoding is used as an intermediate representation between the textual source and the semantic network. The user query is processed by a set of heuristics in a DWIM fashion, so that the requested fields are automatically guessed in most cases. It is then transformed into a semantic graph template and eventually compiled as a Prolog goal, which is executed yielding the search results. The Web UI is based on the PWP suite (Prolog Well-formed Pages). The application code, the data and the UI elements are all packed together into a single portable executable file.

There are two search options available: simple and advanced. A simple search is performed:

1. On request (word, word component, year of approximate appearance of a word in Russian).

2. Alphabetically. There is a search in the Latin alphabet and numbers, since the dictionaries include neologisms that consist of numbers and letters, as well

as neologisms with foreign-language components (initial and final).

Advanced search is possible by the following parameters:

1. Labels (grammatical, stylistic, emotional; labels that indicate the language of borrowing). In the series of annual dictionaries of the current decade, thematic ones have been added to the listed labels.

2. Full-text search on request.

3. By chronological parameter. Temporal boundaries make it possible to find new words of a certain period.

The entire database is built on the material selected by lexicographers manually, which means that the new words are attributed by the time parameter, as well as new meanings, new morphs (affixoids), and new collocations.

The inclusion of materials into the database is preceded by long preliminary work carried out by a large number of professional researchers: published editions of dictionaries are marked up in a certain way; semantic disambiguation is removed; reference entries included in compounds and collocations are duplicated, which makes it possible to remove the problem of formal, meaningless references; to facilitate the search by time parameter, the year is set for each quotation and for each colocation; technical errors are removed.

The *Russian Academic Neography* resource currently does not take into account the usage parameter, since the dictionaries of this series rely on the non-linguistic *Integrum* corpus of texts, the materials of which, however, allow us to identify the number and dynamics of new words used in Russian texts from the mid-1980s until now. In comparison, in the German dictionary database *das Online-Wortschatz-Informationssystem Deutsch* (https://www.owid.de/), the chronological and frequency parameters are presented in the form of diagrams (see the neologic section of *Neologismenwörterbuch*). The diagrams are available for words which appeared in German texts from the 1990s to 2017). Nevertheless, the authors of the *Russian Academic Neography* use data from a number of Russian resources which will give us the option to incorporate this function later. For example, *The National Corpus of the Russian Language* has a section called "graphics" (http://www.ruscorpora.ru/new/graphic.html), charts built on a chronological-frequency principle in this section are based on the *Google Ngram Viewer* service. The *Google Books Ngram Viewer* online search service, which has its own corpus of Russian-language texts, allows you to search for words and compare their usage from 1800 to 2008.

Thus, the *Russian Academic Neography* resource is a set of lexical and phraseological units, reflecting changes in the Russian language over the past 60 years.

5. The scientific potential of the resource

The *Russian Academic Neography* information retrieval resource is intended not only for specialists in the field of Russian lexicology and lexicography, but for all linguists and the wider audience.

Thanks to the query system, the following data is going to be available:

- materials of all new-word dictionaries published since 1971, which are currently a bibliographic rarity;
- the lexical materials of the Russian language (1960–2020s), not recorded in other dictionaries;
- when requesting chronology, it becomes possible to establish the occurrence of a word in a particular period;
- when requesting derivational formants, it becomes possible to identify relevant derivational models;
- with the root query, it becomes possible to identify word-building nests and derivational schemes;
- when requesting a label, the trends of the functional and stylistic dynamics of the vocabulary of the Russian language are identified;
- when requesting a source language, it is possible to reveal all borrowed lexemes in one or another period of time, etc.

Due to the fact that the portal database contains about 116,000 professionally collected and processed new units of the Russian language, which is as many as the average vocabulary of the Russian language represented in explanatory dictionaries, the scientific potential of the *Russian Academic Neography* information retrieval resource cannot be overestimated. Introducing a huge lexical layer of the modern Russian language and the newest Russian vocabulary, it is expected be of great interest to professional linguists and a wider audience.

6. References

Butseva, T. N. (2013). Neologicheskaya sluzhba russkogo yazyka [Neology Service of the Russian Language]. In V. P. Zakharov & M. N. Priemysheva (eds.) *Leksikologiya, leksikografiya i korpusnaya lingvistika.* St. Petersburg: Nestor-Istoriya, pp. 93–98.

Dmitriev, D.V. (2013). Neologia.ru: principy postroeniya internet-resursa dlya kollektivnoj leksikograficheskoj raboty [Neologia.ru: Principles of the Internet

Resource Construction for Joint Lexicographic Work]. In V. P. Zakharov & M. N. Priemysheva (eds.) *Leksikologiya, leksikografiya i korpusnaya lingvistika*. St. Petersburg: Nestor-Istoriya, pp. 99–109.

Kotelova, N. Z. (2015). Teoreticheskie aspekty opisaniya neologizmov [Theoretical aspects of the new words description]. In Kotelova N.Z. *Izbrannye raboty*. St. Petersburg: Nestor-Istoriya, pp. 254–269.

Lesnikov, S. V. (2019a). Akademicheskie tolkovye slovari russkogo yazyka kak yadro akademicheskogo slovarnogo korpusa russkogo yazyka [Academic Explanatory Dictionaries as a Core of the Academic Corpus of the Russian Language Vocabulary]. In *Sbornik nauchnyh statej po itogam raboty Mezhdunarodnogo nauchnogo foruma "Nauka i innovacii: sovremennye koncepcii" (g. Moskva, 5 aprelya 2019 g.)*. Part 1. Moscow: Infiniti, pp. 38–47.

Lesnikov, S. V. (2019b). Akademicheskij slovarnyj korpus russkogo yazyka [Academic Corpus of the Russian Language Vocabulary]. In: *XLVIII Mezhdunarodnaya filologicheskaya nauchnaya konferenciya SPbGU, 18–27 marta 2019* (http://conference-spbu.ru/conference/40/reports/9649).