ISSN 2181-922X

UZBEKISTAN LANGUAGE & CULTURE

O‘ZBEKISTON TIL VA MADANIYAT

KOMPYUTER LINGVISTIKASI

2023 Vol. 1 (6)

www.compling.tsuull.uz

MUNDARIJA

ISSN 2181-922X 18-26

FUNCTIONALITY OF THE RUSSIAN NATIONAL CORPUS

Victor Zakharov[1]

Abstract.

The paper describes the state of the art of Russian corpus linguistics. The main attention is paid to the Russian National Corpus and irs functionality.

Keywords: *Russian corpus linguistics, corpora, the Russian language, the Russian national Corpus.*

1 Introduction

In recent years creation of different text corpora became one of the cutting edge directions in the applied linguistics. In 1980s the Computer Fund of the Russian Language project started. The idea belonged to the academician Andrei Yershov. The idea was stated as follows: “Any progress in the field of construct-ing models and algorithms will remain a purely academic exercise, unless a most important problem of creating a Computer fund of the Russian language is solved. It is to be hoped that creation of such a Computer fund by linguists, quali-fied for the task, will precede construction of large systems for application pur-poses. This would minimize labor costs and simultaneously would protect the 'tissues' of the Russian language from arbitrary and incompetent intervention” [Yershov, 1979]. Corpora were an integral and crucial part of this project.

The Russian National Corpus (RNC) is the most popular one among lin-guists for both being the most well-known and the opportunities which it pre-sents. However, being unable to go into a deeper analysis within the framework of this paper, we will zero on its general characteristics together with its most unique features.

[1] Victor Zakharov - candidate of philological sciences, associate professor, Saint-Petersburg State University, Saint-Petersburg, Russia.
E-mail: v.zakharov@spbu.ru
ORCID: 0000-0003-0522-7469

2 The Structure of the RNC

The Russian National Corpus (http:// ruscorpora.ru) includes original prose, translations (parallel with original texts), poetry, as well as texts, repre-senting the non-standard forms of modern Russian [Natsionalniy korpus russ-kogo yazyka, 2005, 2009]. It was started in 2003 and from 2004 is accessible via Internet. The corpus size in total is about 1500 million tokens (March 2023).

The corpus allows us to study the variability and volatility of linguistic phenomena frequencies, as well as to obtain reliable results in different areas: the study of morphological variants of words and their evolution; the study of chang-es in lexicon; syntactic relations; the research of changes in Russian accent; se-mantic relations; and so on.

The RNC is a collection of individual corpora, each of which is assembled to tackle different linguistic tasks. Each of these collections of texts is large and representative, which makes them valuable for both quantitative and qualitative research. The specificity of linguistic tasks determines the structure of each corpus and the type of annotation used in it.

Within the main corpus the RNC includes the following subcorpora:

1) The Main corpus. The main corpus is the most general reference corpus. It comprises prosaic texts created after 1700, printed, written, or (later) electron-ic. The main corpus counts in total 373 million tokens. The part of modern texts is the largest one. Texts are represented in proportion to their share in real-life usage. For example, the share of fiction does not exceed 40%.

Every text included in the main corpus is subject to meta-tagging and mor-phological tagging. Morphological tagging is carried out automatically. In a small part of the main corpus (around 6 mln tokens) grammatical homonyms are dis-ambiguated by hand, and results of automated morphological analysis are cor-rected. This part is the model morphological corpus and serves as a testing ground for various search algorithms and programs of morphological analysis and automated processing. Disambiguated texts are automatically supplied with indicators of stress. Stress annotation may be turned off for printing or saving the search results.

2) The Corpus of Spoken Russian. It represents real-life Russian speech and includes the recordings of public and

spontaneous spoken Russian and the transcripts of the Russian movies. To record the spoken specimens the standard spelling was used. The corpus contains the patterns of different genres/types and of different geographic origins. See more in Sect. 4.

3) Deeply Annotated Corpus (SynTagRus treebank). This corpus contains texts augmented with morphosyntactic annotation. Besides the morphological information, every sentence has its syntax structure (disambiguated). The corpus uses dependency trees as its annotation formalism (Fig. 1).

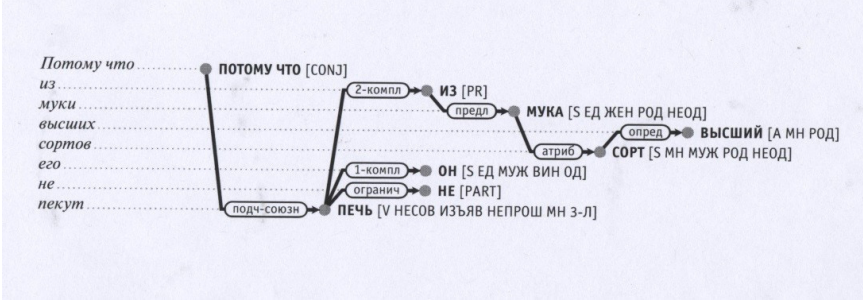

Fig. 1. Dependency tree of a Russian sentence

Nodes in such a tree are words of the sentence, while its edges are labeled with names of syntax relationships.

4) Parallel text corpus. The site contains parallel texts of 26 language pairs.

5) Dialectal corpus. The dialectal corpus contains recordings of dialectal speech (presented in loosely standardized orthography) from different regions of Russia. The corpus employs special tags for specifically dialectal morphological features; moreover, purely dialectal lexemes are supplied with commentary.

6) Poetry corpus. At the moment the poetry corpus covers the time frame between 1750 and nowdays. Apart from the usual morphological tagging, there is a number of tags adapted for poetry.

7) Accentological corpus is a primary tool for studying the Russian word stress, which is a key feature of its annotation.

8) Educational corpus. The educational corpus is a small disambiguated corpus adapted for the Russian school educational program.

9) Media (newspaper) corpus. It is the largest one within the RNC. It con-sists of media publications that have appeared since 1980s.

10) Multimodal/multimedia corpus. It consists of

synchronized video or audio recordings, for several films, where text, sound and gestures are annotated. See more in Sect. 4.

11) Corpus "From 2 to 15". It is dedicated to the texts read by children and teenagers, automatically annotated by the presumed age of their audience.

12) Russian classics. This corpus includes texts of classical Russian litera-ture,

13) Historical corpora. There are several historical corpora within the RNC: texts representing the Old East Slavic language (from the 11th to the 14th centu-ry), Middle Russian (the language of the 15th to the 17th centuries) and Old Church Slavonic in its Russian version. In another historical corpus, birch-bark letters of the 11th-15th centuries are collected.

14) Panchronic corpus. This is a united search service in the historical and contemporary corpus, which allows to trace the history of a given word or grammatical construction throughout several centuries.

3 Semantic annotation in the Russian National Corpus

Semantic annotation is a unique feature of RNC that makes it distinct from other national corpora. There are three groups of tags assigned to words to reflect lexical and semantic information: class, lexical and semantic features, derivational features. The set of semantic and lexical parameters is different for different parts of speech. Moreover, nouns are divided into three subclasses (concrete. abstract and proper names), each with its own hierarchy of tags. Lexical and semantic tags are grouped as follows: taxonomy, mereology, topology, causation, auxiliary status, evaluation [Laveshskaja, Shemanaeva, 2008].

The meta-language of tags is based on English notation; it is, however, pos-sible to make a search using traditional Russian category names in the search “semantic features” form. The following are some tags from an inventory of available tags with examples in parenthesis. Some tags for concrete nouns:

Taxonomy: t:hum – person (*человек* (human), *учитель*(teacher)), t:hum:etn – ethnonyms (*эфиоп* (Ethiopian), *итальянка* (Italian)), t:hum:kin – kinship terms (*брат* (brother), *бабушка* (grandmother)), t:animal – animals (*корова* (cow), *сорока* (magpie)), etc.

Some tags for verbs:

t:move – movement (*бежать* (run), *бросить* (throw))

t:put – placement (*положить* (put), *спрятать* (hide))

t:impact – physical impact (*бить* (beat), *колоть* (prick))

t:be:exist – existence (*жить* (live), *происходить* (happen))

t:be:appear – start of existence (*возникнуть* (arise), *создать* (create))

t:be:disapp – end of existence (*убить* (kill), *улетучиться* (diappear))

t:loc – location (*лежать* (lie), *стоять* (stand)).

These are just a few examples of 200 tags available in the corpus that are structured in a hierarchical way.

4 Speech corpora of Russian

Oral speech, and especially, the nonpublic oral improvised speech is the most important version of language. Therefore it is important to dwell on Russian speech corpora.

The Corpus of Spoken Russian is the collections of transcripts of the spo-ken texts of different types. Its volume just now is around 13,4 million tokens. These transcripts are annotated morphologically and semantically by the RNC annotation system. In addition, the corpus has its own annotation: the accento-logical and the sociological one. The sociological annotation means that to every text the information on the sex, the age is assigned, so a user can form his own subcorpora according to all these parameters and their combinations.

The Spoken corpus of the RNC gives a user various possibilities, but all these tasks must not be connected or based on the real phonation. Therefore, ***the Multimedia Russian corpus (MURCO)*** was formed as a part of the RNC [Grishi-na, 2009]. Its material are fragments of movies of the 1930s through the 2000s. The main principle of the MURCO is the alignment of the text transcripts with the parallel sound and video tracks. Consequently, when a user makes his query he may obtain not only a written text, annotated from different points of view, but also the corresponding sound and video material. The total volume of the movie transcripts in the RNC is around 5,7 million tokens.

The types of specific annotation in the MURCO are as follows: a) orthoepic annotation: combinations of sounds are marked; b) annotation of accentological structure: the word structure in regard to the stress position is defined; c) speech act annotation: the types of speech acts and vocal gestures, used in a clip are de-scribed; d) gesture annotation: the type of gesticulation in a clip is described.

5 Search in the Russian National Corpus

One can search by an exact form, by a set phrase, by lexico-grammatical and semantic features, by additional features such as a specified position (before or after punctuation marks, in the beginning or in the end of a sentence, capitali-zation, etc). Words we are searching for could be combined with logical operators «AND», «OR» and «NOT». It can be used with both left or right truncation. Dis-tance between words could be set from minimum to maximum. The distance be-tween words next to each other is 1 word. For lexico-grammatical search, we can input a sequence of lexemes and/or word-forms with certain grammatical and/or semantic features. We can combine them in any way.

A simpler way to search for certain grammatical features is to use a selec-tion window. The selection window contains a list of appropriate features, subdi-vided by categories: i.e., part of speech, case, gender, voice, number for morphol-ogy, etc.

The ***semantic features*** field allows for listing the semantic and derivational features of the lexeme. As a rule, semantic features have a hierarchy. *By default* all the tagged meanings of a given word are searchable. For instance, the parame-ter Human qualities selected in the Semantic features field will yield both *умный* 'intelligent', *верный* 'faithful', *коварный* 'perfidious' (where the parameter is present in its basic meaning), as well as *мягкий* 'soft' or *холодный* 'cold' that apply to human beings only metaphorically. To refine the scope of the search, we could select one or two parameters: "sem" — only the first meaning given in dictionar-ies is searched (thus human qualities will yield words like 'intelligent', 'faithful' or 'perfidious', but not those like 'soft' or 'cold'); "sem2" — the meanings other than the first ones are searched (thus only words like 'soft' or 'cold' will be found).

The types of annotation which are specific for the special corpora (MUR-CO, poetical, dialectological, etc.) define the peculiarities of the appropriate inter-face in comparison with the interface of the RNC proper.

Search results can be presented twofold: a horizontal text (a broader con-text) and a concordance (Fig. 2). In both cases grammatical and semantic features of any word can be checked out (Fig. 2 shows that for the word *женщина* (wom-en)).

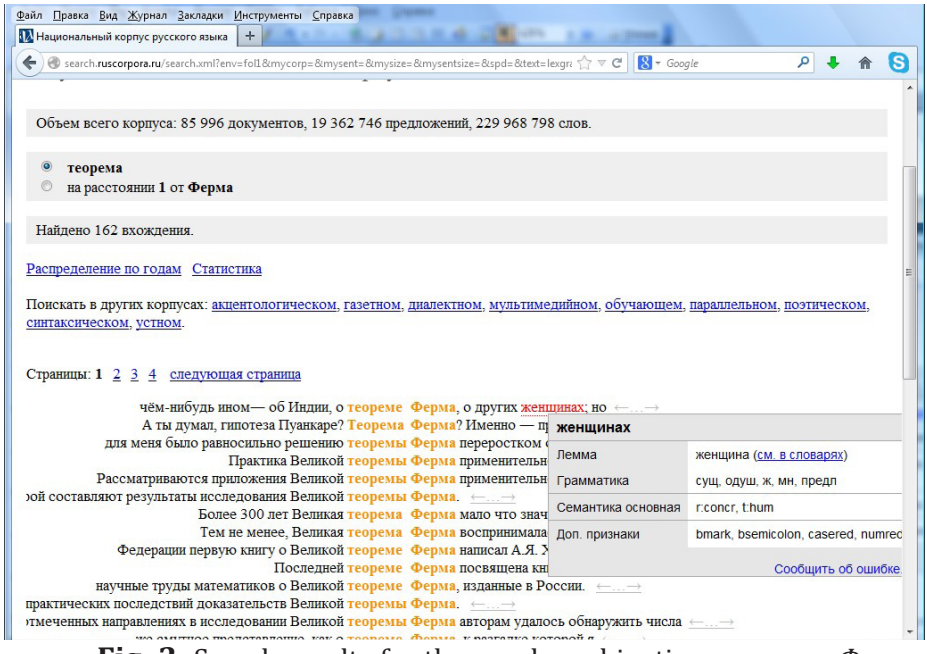

Fig. 2. Search results for the word combination *теорема Ферма* (Fermat's theorem)

From the search page one can get to other forms of result representation (graph by year, statistics, frequency, n-gramms).

6 New Features

In 2022-2023, the RNC has gone through a major modification phase. The website has been redesigned. The start page and the pages with general infor-mation on the corpus are now displayed with a new interface. The project de-scription has been revised and updated. A FAQ section is added explaining the main features of the RNC. The English version of the site has also been partially updated. The new website is fully adapted for mobile devices. Gradually, all of subcorpora will switch to the new interface.

The main corpus was updated by new texts of different genres and types including a collection of different academic genres (abstracts, programs, text-books, problems), a collection of technical guides and instructions end so on.

More search results could be downloaded in Excel format from the main and media corpora. As many as 5000 examples can be saved into an Excel table.

The syntax corpus has been significantly updated with information about the texts, namely the gender of the author, the topic and type of the text, its source.

An important innovation in the parallel corpus is that the

original and translated texts are now shown in two columns.

The function «Similar words» appeared. These are the words that are se-mantically closely related to the word in question and are used in similar contexts.

The corpus of regional media is searchable for collocations. For this search mode a statistical approach is used. Collocations are combinations of words that occur together more often than by chance. Such statistical measures as Dice, Log-likelihood, t-score, MI3 and aggregated measure are used to calculate the collocations.

Each corpus within the RNC has its own Corpus Portrait. The Corpus Por-trait functionality is designed as a tool that allows a RNC user to analyse the characteristics of a given corpus and assess whether the corpus in question is suitable for their research or teaching needs. All the RNC corpora have tags on a metacorpus level, allowing to categorize them by historical period, text type, presence of specific annotation, etc.

To see all the information on a given word, ont can now use the Word at a glance functionality. As of today, the Word Portrait includes: grammatical and semantic properties of the word; similar words (only in the main corpus); word usage examples in the corpus; distribution of examples by year and by type of text; the compatibility (collocations) with words of different parts of speech.

See this information in more detail in [Savchuk, 2021] and on the corpus site in the section RNC News (https://ruscorpora.ru/en/news).

7 Conclusion.

Now all corpora of the Russian language and mostly the RNC are used by both Russian and foreign researchers. The RNC has English interface and the help system in English. Its subcorpora with their special annotation provide many possibilities for linguistic studies.

The studies based upon the semantic annotation are of special interest. There are works which address word sense disambiguation and lexical constructions. The type and degree of specification of the RNC semantic annotation and new tools developed in the last years enable various semantical and statistical re-searches.

References

Yershov A.P. K metodologii postroeniya dialogovykh sistem. Fenomen delovoi prozy. Novosibirsk (1979).

Natsionalnyi korpus russkogo yazyka: 2003–2005. Moskva (2005).

Natsionalnyi korpus russkogo yazyka: 2006 – 2008. Saint-Petersburg (2009).

Lashevskaja O.N., Shemanaeva O.Ju. Semantic Annotation Layer in Russian Na-tional Corpus: Lexical Classes of Nouns and Adjectives In: Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, Morocco (2008).

Grishina E. Multimodal Russian Corpus (MURCO): General Structure and User Interface. In: Slovko 2009. NLP, Corpus Linguistics, Corpus Based Gram-mar Research. Jana Levická, Radovan Garabík (eds.). Bratislava, Slovakia, pp. 119-131 (2009).

Savchuk S.O. National corpus of the Russian language in the mirror of statistics // Proceedings of the international conference "Corpus Linguistics-2021". St. Petersburg, 2021. pp. 18-30 (2021).