Russian Language Studies
Русистика
ISSN 2618-8163 (Print); ISSN 2618-8171 (Online)
2024 Vol. 22 No. 4 501–517
http://journals.rudn.ru/russian-language-studies

DOI: 10.22363/2618-8163-2024-22-4-501-517
EDN: AMYSNF

Introductory article

Approaches and tools for Russian text linguistic profiling

Marina I. Solnyshkina[1] ✉, **Valery D. Solovyev**[1], **Yulia N. Ebzeeva**[2]

[1]Kazan (Volga Region) Federal University, *Kazan, Russian Federation*
[2]RUDN University, *Moscow, Russian Federation*

✉ mesoln@yandex.ru

Abstract. Approaches and tools for assessing linguistic and cognitive complexity of educational texts are in demand both in science and teaching. Predicting difficulties of perception and understanding and ranking texts by classes, i.e. the number of years of learning or levels of language proficiency (A1–C2), are of particular importance for education. The study is aimed at demonstrating modern methodologies, algorithms, and tools for analyzing Russian texts in text profiler and automatic analyzer RuLingva and at presenting articles from the thematic issue on comprehensive analysis of Russian language textbooks for Russian and Belarusian schools. The research demonstrates that the modern paradigm of discourse complexology is based on the methods of stylistic statistics, which identifies functional characteristics of language units and verifies them using big data. The services on RuLingva are designed for teachers and researchers; they automatically analyze educational texts and predict their target audience based on readability, lexical diversity, abstractness, frequency, and terminological density. In "Russian as a Foreign Language" mode, RuLingva downloads lists of words from the text according to each level of language proficiency and estimates their proportion. This provides material for pre- and post-text work. RuLingva algorithm is based on the typology of educational texts and is to be supplied with tools for assessing a person's verbal intelligence and reading literacy. The nearest prospect of RuLingva lies in widening the range of complexity predictors and installing automatic subject area discriminator. Both directions are planned to be implemented using neural networks, classification models, "typological passports" of educational texts with different complexity, and thematic orientation.

Keywords: linguistic analysis, text profiler RuLingva, text complexity, educational text, typological passport of the text, complexity predictors

Contribution: Solnyshkina M.I. — idea, research, text preparation and editing; Solovyev V.D. — methodology, research; Ebzeeva Yu.N. — research, approval of the final version of the article.

Funding. This article has been supported by the Kazan Federal University Strategic Academic Leadership Program (PRIORITY–2030). This publication has been supported by the RUDN University Scientific Projects Grant System, project no. 050738-0-000.

Conflict of interests. The authors declare that they have no conflict of interests.

Article history: received 02.07.2024; accepted 18.08.2024.

For citation: Solnyshkina, M.I., Solovyev, V.D., & Ebzeeva, Y.N. (2024). Approaches and tools for Russian text linguistic profiling. *Russian Language Studies, 22*(4), 501–517. http://doi.org/10.22363/2618-8163-2024-22-4-501-517

Introduction

The change of modern scientific paradigms and active integrative processes have set new tasks for linguists, which imply, on the one hand, the inclusion of text in broad historical and discursive contexts, and on the other hand, the study of the processes of text perception, understanding, reproduction, and generation. The very fact of addressing the discursive aspects of text and the cognitive characteristics of the native speaker expanded the boundaries of linguistics, involved data from other sciences, and substantiated the use of more than one approach to analyze data.

Among the most urgent tasks of text analytics, scholars single out text classification, tone analysis, keyword extraction, "diagnosis" of the types of relations between text units, determination of semantic roles, analysis of arguments and discourse structures, structuring of large linguistic data, etc. (Kuznetsova, 2015; Young et al., 2018). The tasks of homonymy/polysemy resolution (removal), as well as thematic modeling of the text are of particular complexity (Sakhovskiy et al., 2020). One more task is the creation of author linguistic profile with a set of quantitative characteristics peculiar to a particular author (Mikheev, Ehrlich, 2018). These multidimensional tasks imply access to large collections of texts of various forms, registers, types, and genres and the use of automatic analysis tools.

When setting research goals, a scientist chooses an approach and appropriate methods, collects data, and selects appropriate tools. Now we may choose not one but several approaches, including interdisciplinary ones, and use large representative electronic corpora, including those created earlier. A corpus of linguistic data contains not only a meta markup, but also a detailed description of each text, its "typological passport", its quantitative characteristics, its "linguostatistical profile" (Virk et al., 2020). The "profile" contains data on the frequency, distribution of text linguistic parameters, and the relations between text linguistic characteristics. The latter means that texts of different types, genres and registers are "profiled" according to their features and ranges of reference values of these features. Reference values perform predictive and discriminant functions; they determine the text genre, type and register and differentiate texts as elements of certain types, genres and registers. Text profiling and language data matrices are the final stage of corpus collection and organization. Text profiles need general scientific approaches of text analytics and specific algorithms for automating linguistic analysis and text analyzers for automatic evaluation of text parameter values (Lukashevich, Dobrov, 2015; Namestnikov, Pirogova, Filippov,

2021; Solovyev, Solnyshkina, McNamara, 2022; Kolmogorova, Kolmogorova, Kulikova, 2024). Effective automation of text "profiling" and general labor-intensive mechanical tasks of linguistic analysis of texts in Russian bring Russian language studies and Russian linguistics in general to a qualitatively new level.

Modern scholars refer to text "profiles" as so-called "resources" of a language, and languages are divided into high-resource and low-resource languages, depending on the sufficiency of data for machine learning or other types of processing (Chang et al., 2023). Linguistic typology has an analogue contrast between well-described and under-described languages. The former include, for example, English and German. Russian in this respect is qualified as a low-resource (Valeev et al., 2019) or "relatively high-resource" language (Karakanta, Dehdari, van Genabith, 2018). However, it is still necessary to create electronic databases and tools and improve text analytics approaches for Russian (Toldova et al., 2015).

Consequently, principles of linguistic profiling and automation of language data analysis are a relevant issue. **The aim of the study** is to describe theoretical approaches and tools of linguistic profiling of texts in Russian. The second part of the paper presents articles of the thematic issue.

Linguistic profiling in theoretical and applied linguistics

Methods of exact sciences and mathematical models are traditional for text description. The works of F. de Saussure at the beginning of the XXth century (1922, first edition in 1916) (Saussure, 1977) were followed by the interdisciplinary research of C. Shannon and W. Weaver (1949), which laid the foundations of methods of quantitative linguistics. The approach to linguistic phenomena as stereotypical, labeling certain phenomena and characteristics as inherent or alien to some type of objects (Lipmann, 1922), is important. Stereotypes typify texts and identify the parameters peculiar to each type. One of the first hypotheses concerning the statistical differences of discourses belongs to V.V. Vinogradov. In 1938, Vinogradov wrote, "Apparently, different styles of bookish and colloquial speech, different styles and genres of fiction, show different frequency of types of words. Unfortunately, this question is only in the preparatory stage of survey" (Vinogradov, 1938:356). In 1930–1960, Russian and foreign linguistics made great progress, so that the linguistics of the 1960s was called "the most precise of all humanities", primarily due to the clear and formalized theory of N. Chomsky, applicable not only to natural, but also to programming languages.

This is due to universal principles, models of semantic constants (see Krongauz, 2009) and syntactic constructions as the main objects of research. Formal models were created considering linguistic units as components of a linguistic system organized according to universal cognitive principles and linguistic unit functioning in a text of a certain type (Zinder, Stroeva, 1968).

Text analytics for the Russian language was developed by B.N. Golovin (Golovin, 1971) and his scientific school, who widely used quantitative methods to describe and analyze functional styles. According to M.A. Kormilitsyna, O.B. Sirotinina, the main merit of the Gorky consisted in creating the system of statistical methods for studying speech facts. These methods are based on the strong correlation between semantic and distributive properties of linguistic units" (Kormilitsyna, Sirotinina, 2013: 103).

The stylistic-statistical (qualitative-quantitative) method developed in the Russian school at the end of the XXth century is of particular importance. This involves (1) "semantic-stylistic qualification" of linguistic units, i.e. revealing their specific functional characteristics, and (2) verification of these characteristics with mathematical statistics methods (see Kozhina, 1989). This method, which became popular in the 1980s due to the development of formal language, is also used in modern quantitative linguistics to assess the influence of a factor on a construct (Serdobolskaya, Toldova, 2005).

The turn of the millennium saw the emergence of computer and corpus linguistics (Solnyshkina et al., 2022) and numerous approaches to formalized processing of large linguistic data. Text analysis changed significantly due to the revolution in computational technologies and databases. Quantitative methodologies, including machine learning, made the extraction of information from text data arrays accessible, and consequently, allowed us to approach the confirmation/refutation of earlier hypotheses about the systematicity of linguistic facts in texts of certain genres, registers, and types.

These changes in science are complementary stages developing models of three types: feature-based, representation learning, and generative models. Modern automatic Russian language text analyzers continue to gain popularity. At this stage, open platform solutions are offered by the text profiler Textometr, which is actively used by Russian word processors (Laposhina, Lebedeva, 2021), and I. Begtin's analyzer "Text Readability Assessment" (I. Begtin, 2021)[1], which has 5 readability formulas. I. Begtin's project became the first online server with built-in readability formulas, but, unfortunately, the developers suggest to consult English-language Wikipedia sites[2] to study the calculation algorithm and formulas. This generally does not allow us to assess the validity of the formulas.

RuLingva text profiler and text complexity analyzer was developed within the framework of the Russian Science Foundation project "Complexity of Texts in Russian"[3]. The project has two main goals: to identify and describe typological

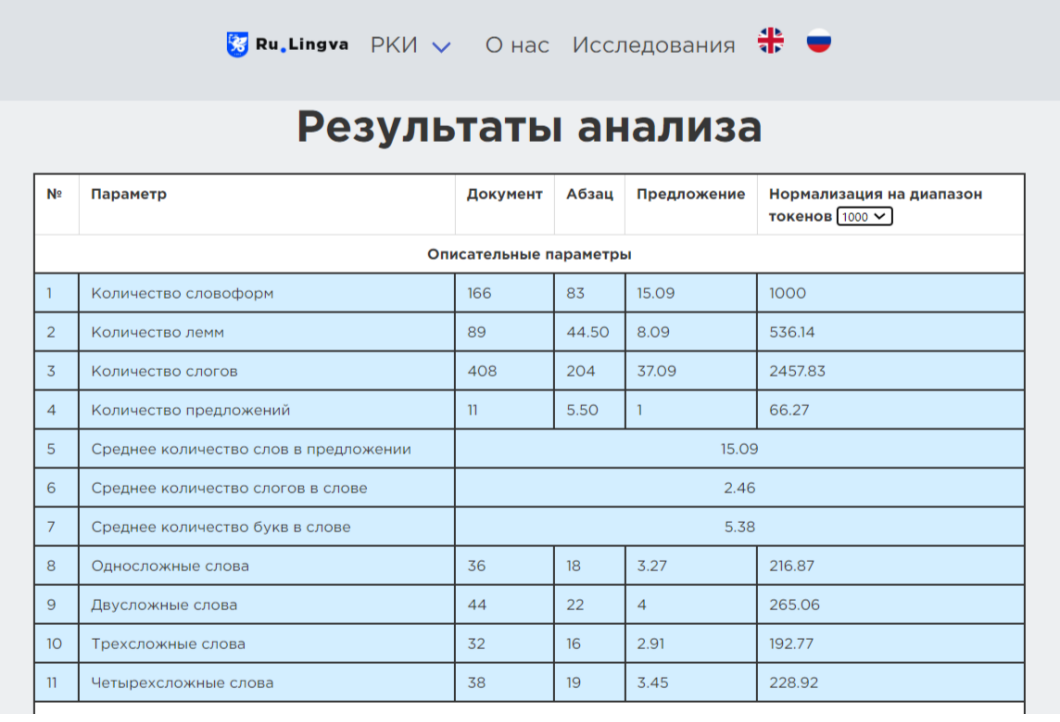
[1] A convenient tool for assessing texts. Retrieved June 02, 2024, from https://plainrussian.ru/#about (accessed on 02.06.2024).

[2] Flesch – Kincaid readability tests. Retrieved June 16, 2024, from https://en.wikipedia.org/wiki/Flesch%E2%80%93Kincaid_readability_tests Retrieved June 16, 2024, from https://en.wikipedia.org/wiki/Coleman%E2%80%93Liau_index

[3] Card of the project of fundamental and exploratory scientific research, supported by the Russian Science Foundation. Retrieved June 18, 2024, from https://rscf.ru/prjcard_int?18-18-00436

parameters of academic texts and to develop methods for ranking texts by complexity levels. The complexity level of texts in RuLingva is evaluated according to correlations between the parameters of texts and readers' features (age, education, vocabulary, etc.).

Following the modern tradition in text analytics, we use the terms *text characteristics*, *text parameters*, *values*, or *metrics*, and clusters or groups of parameters (which D. McNamara (McNamara et al., 2014) refers to as brands). The term *characteristic* denotes the name of a linguistic category (e.g., lexical diversity); the term *parameter* provides information about the way(s) to evaluate the relevant text characteristics. For example, lexical diversity (a text characteristic) is estimated through Type Token Ratio (TTR) parameter, i.e. the ratio of the number of word forms to lemmas. The terms *metrics* and *values* are interchangeable and show quantitative values of the parameter. For example, value/metric 166 in row 1 (fig. 1) indicates the number of word forms (parameter) which shows text length (feature). Conceptually similar text parameters are grouped into clusters. For example, a descriptive cluster of text parameters includes text length measured in the number of word forms, lemmas, syllables, or sentences.



Результаты анализа

№	Параметр	Документ	Абзац	Предложение	Нормализация на диапазон токенов 1000
	Описательные параметры				
1	Количество словоформ	166	83	15.09	1000
2	Количество лемм	89	44.50	8.09	536.14
3	Количество слогов	408	204	37.09	2457.83
4	Количество предложений	11	5.50	1	66.27
5	Среднее количество слов в предложении	15.09			
6	Среднее количество слогов в слове	2.46			
7	Среднее количество букв в слове	5.38			
8	Односложные слова	36	18	3.27	216.87
9	Двусложные слова	44	22	4	265.06
10	Трехсложные слова	32	16	2.91	192.77
11	Четырехсложные слова	38	19	3.45	228.92

Fig. 1. RuLingva Interface
S o u r c e : RuLingva. Retrieved June 18, 2024, from https://rulingva.kpfu.ru/

The RuLingva profiler[4] supported by the research group of Kazan Federal University calculates the values of 73 parameters of Russian educational texts.

[4] RuLingva. Retrieved June 18, 2024, from https://rulingva.kpfu.ru/

According to the modern tradition in computational linguistics, linguistic parameter values are measured with varying degrees of *granularity* (see the term in Paraschiv et al., 2023), i.e. metrics are calculated within a sentence, paragraph, text fragment of a certain length, and the whole document. The user can set the metrics calculation normalization depending on the research tasks at 100, 200, or 1000 word forms (tokens) (see fig. 1).

Before developing RuLingva functionality, two independent corpora were created: the Educational Corpus of the Russian Language (hereinafter referred to as ECRL) and the Corpus of Russian as a foreign language texts (hereinafter referred to as CRFLT). At this stage the volume of the ECRL[5] is 14 million word-forms; the volume of the CRFLT is a little more than 500, 000 word-forms. The underlying principle in creating both corpora was the principle of data reliability, so the corpus included only "reference" texts, i.e. texts that had undergone professional expertise and were recognized as the best texts in their field.

The sources of materials for the ECRL were the texts of the Federal State Educational Standard[6]; the texts for CRFLT were chosen from texts recommended by the Commission for the Examination of Test Materials in Russian as a Foreign Language[7], the Expert Commission of the State System of Testing Foreign Citizens in Russian[8], and texts from the Open Assignment Bank of the Federal Institute of Pedagogical Measurement[9]. The differential completeness, the balance and representativeness of the ECRL, which was used as a source in domestic and foreign studies, is beyond doubt (Corlatescu et al., 2022, Kupriyanov et al., 2022, Paraschiv et al., 2023). This proves that RuLingva is a valuable source for studying modern scientific and academic discourse and profiling Russian texts.

The ECRL and CRFLT corpora are closed and used only for research purposes. A small demo sample is publicly available. It is a part of a sub-corpus of educational texts of the subject block, which includes random texts from Russian social studies textbooks (CORAT, Corpus of Russian Academic Texts[10]). To retain copyright, the sequence of paragraphs and sentences in CORAT texts has been changed.

The first formula of readability of Russian educational texts is based on the Russian Language Learning Corpus:

[5] Certificate of state registration of the database № 2020622254.

[6] Federal list of textbooks. Retrieved June 18, 2024, from https://fpu.edu.ru/

[7] Order "On approval of the Regulation on the Commission for the examination of test materials in Russian as a foreign language and its composition" Retrieved June 18, 2024, from https://docs.cntd.ru/document/901860364

[8] Order of February 16, 2005 No. 69 "On the establishment of an expert commission of the state system for testing citizens of foreign countries in the Russian language" Retrieved June 18, 2024, from https://normativ.kontur.ru/document?moduleId=1&documentId=85661

[9]Exam for foreign citizens and stateless persons. Retrieved June 18, 2024, from https://fipi.ru/inostr-exam

[10] Research Laboratory "Multidisciplinary Text Research" Retrieved June 18, 2024, from https://ifmk.kpfu.ru/laboratory/tekstovaya-analitika/

Flesch — Kincaid Index (SIS) = 208.7 – 2.6 × ASL – 39.2 × ASW, where ASL is average sentence length, and ASW is average word length in syllables (Solovyev et al., 2018). After successful validation on humanitarian, philological and natural science texts of subject blocks for middle and high school (Gatiyatullina et al., 2020), the formula was installed on the RuLingva website and is used to assess the readability of educational Russian texts[11]. This formula is convenient because it ranks the readability of educational texts by years of schooling, i.e., grades. For example, a text with a 7.62 readability (Flesch — Kincaid Index (SIS)) is for grades 7–8 (fig. 2).

To assess the fiction prose texts readability on RuLingva site, the Flesch-Kincaid readability formula was modified by I.V. Oborneva for the Russian language:

Flesch — Kincaid index (O) = 206.835 – 1.3 × ASL – 60.1 × ASW.

I.V. Oborneva defined this formula on the materials of the author's English-Russian corpus of parallel fiction texts, so it is recommended only for assessing the readability of fiction prose texts (Oborneva, 2006). I.V. Oborneva's formula gives higher results when assessing the readability of educational texts (fig. 2) (Kupriyanov et al., 2022).

Параметры читабельности		
12	Индекс Флеша-Кинкейда (SIS)	7.62
13	Индекс Флеша-Кинкейда (O)	12.60

Fig. 2. Text readability parameters on RuLingva
S o u r c e : RuLingva. Retrieved June 18, 2024, from https://rulingva.kpfu.ru/

In addition to readability indices, RuLingva calculates values of four groups of parameters: (a) descriptive (number of words, sentences, syllables, lemmas, and word forms); (b) morphological (number of different parts of speech and their categories); (c) lexical (frequency, abstractness, number of terms of seven subject areas, including philology, mathematics, computer science, natural science, physics, fine arts, music, as well as the number of unique, i.e. non-repeatable words); 4) discursive (local and global word repetitions).

RuLingva evaluates the level of lexical diversity (TTR) of a text, measuring the degree of specificity/abstractness, frequency, and lexical density. Automated lexical diversity value estimation, despite its apparent simplicity, requires a special approach. The calculations of this parameter are reliable only for fragments of 200 to 1000 words (Cvrček, Chlumská, 2015), since the high proportion of service parts of speech in longer texts significantly reduces this parameter. That is why RuLingva automatically divides texts into 1000-word fragments, and average lexical diversity value of the whole document is based on the data about each of the fragments.

[11] RuLingva. Retrieved June 18, 2024, from https://rulingva.kpfu.ru/, RuLex. Retrieved June 18, 2024, from https://rulex.kpfu.ru/nlp

The text abstractness/concreteness index on RuLingva is calculated based on the Russian Foundation for Basic Research project data (Solovyev et al., 2022)[12]. The abstractness/concreteness data were generated from experimental data of Internet crowdsourcing among native speakers, and later three versions of the abstract words dictionary were created: (1) a dictionary of 22,000 words, built on deep learning technology based on the BERT model; (2) a dictionary of 64 thousand words, built on the word2vec technology; (3) a dictionary of 88 thousand word forms, based on the Google Books Ngram corpus (Solovyev et al, 2022).

For Russian as a foreign language texts on the Rulingva website, the shares of vocabulary from A1 to C2 are calculated, as well as the share of words missing in the lexical minima (fig. 3).

53	Доля слов уровня A1	96	48	8.73	57.83
54	Доля слов уровня A2	21	10.50	1.91	12.65
55	Доля слов уровня B1	14	7	1.27	8.43
56	Доля слов уровня B2	27	13.50	2.45	16.27
57	Доля слов уровня C1	4	2	0.36	2.41
58	Доля слов уровня C2	0	0	0	0

Fig. 3. Lexical analysis of a Russian text for foreign students on RuLingva
Source: RuLingva. Retrieved June 18, 2024, from https://rulingva.kpfu.ru/

RuLingva offers data on lexical frequency (fig. 4), classifying all words in the document into groups from A1 to C2 based on their frequency in the Russian National Corpus (Lyashevskaya, Sharov, 2009). The service offers data on the proportion of words of each level, as well as words missing in lexical minima, and allows uploading word lists, giving the teacher material for pre- and post-textual work.

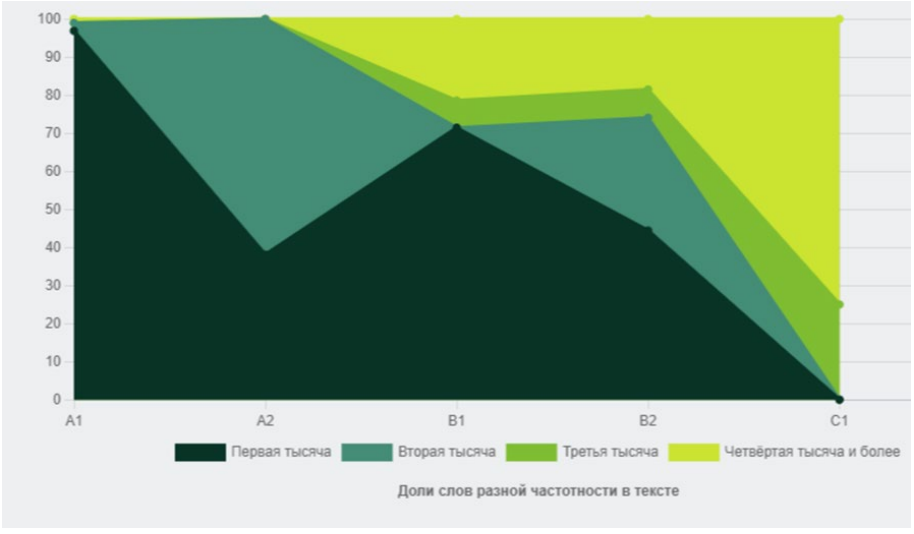


Fig. 4. Frequency analysis of vocabulary in a Russian text for foreign students on RuLingva
Source: RuLingva. Retrieved June 18, 2024, from https://rulingva.kpfu.ru/

[12] OpenLab "Quantitative Linguistics" Retrieved June 17, 2024, from https://kpfu.ru/tehnologiya-sozdaniya-semanticheskih-elektronnyh.html

For researchers aiming at analyzing large amounts of data, RuLingva offers batch processing that allows loading several files for parallel analysis. The report is uploaded with a detailed description of the results of the analytical process in Excel spreadsheet format.

The predictive power of the presented parameters as level of complexity predictors and subject domain discriminants has been proven in a number of studies (Laposhina et al., 2019; Blinova, Tarasov, 2022; Dmitrieva, Laposhina, Lebedeva, 2021; Morozov, Glazkova, Iomdin, 2022; Lyashevskaya, Panteleeva, Vinogradova, 2021).

According to the modern quantitative linguistics paradigm, the research algorithm on the platform includes the following stages:

1. Corpus preprocessing, involving standardized procedures of removing signs of other semiotic systems from the text to preserve the integrity and purity of the input text.

2. Creation of a matrix of parameter values of the analyzed, i.e. uploaded to RuLingva, texts and its subsequent upload in an Excel table.

3. Calculation of average values of each parameter and identification of reference ranges, i.e. variables characteristics of the investigated text.

4. Generalization and identification of universal statistically significant patterns.

The review of the thematic issue

The issue includes papers devoted to Russian language and literature textbooks. Two articles compare modern and Soviet textbooks.

The opening article *Predicative potential of lexical parameters: text complexity assessment in Russian language textbooks for 5–7 grades* by Mariia I. Andreeva, Radif R. Zamaletdinov, Anna S. Borisova considers the linguistic parameters of the text complexity. The first part of the article describes in detail the methodology of creating the necessary corpus of textbooks. It is important that the authors could select the line of textbooks for different grades by the same author. An essential stage of corpus creation is text preprocessing: lemmatization, segmentation, etc. This part of the paper can be useful for all researchers who create text corpuses to study complexity. Further, the authors use two text profilers, RuLingva for estimating the values of 49 language parameters and RuLex for extracting terms from textbook texts. Nine parameters with statistically significant complexity correlation are identified. Interestingly, there was no TTR, a parameter characterizing lexical diversity of the text, among them. The article results in establishing the relationship between text complexity and lexical density (the share of main parts of speech) and text cohesion (the number of lexical repetitions). For the first time, the authors studied the number of terms in textbooks for different grades. The unexpected result was that there are more terms in textbooks for the 5th grade. This result requires further research and discussion. This is the first de-

tailed study of linguistic parameters of the complexity of Russian language textbooks for different grades.

The article by E.N. Bulina, M.I. Solnyshkina, and Y.N. Ebzeyeva *Russian language textbook as agent of change: from USSR to the new century* studies the structure and typography of textbooks of 1935–1974 and 2012–2015. The authors show that the main structural elements of textbooks of different periods coincide and are approximately similarly arranged; they include texts on theory, texts of instructions and tasks, and texts of exercises. However, the share of these three "formants" varies significantly. The volume of tasks in modern textbooks is more than doubled. The nature of instructions has changed. Instructions in Soviet textbooks have the traditional form of inducement expressed by verb imperatives. At the same time, motivational questions prevail in modern textbooks because of the tendency to dialogicality. The authors of the article scrutinize the typography of textbooks. The typography of modern textbooks is more diverse and qualitative, which contributes to a better perception of the text. The article characterizes carefully selected textbooks. The texts are analyzed with modern computer linguistics tools, including the RuLingva software package developed at Kazan Federal University. Considerable attention is paid to general pedagogical issues in the context of the changing socio-political situation in the country. This article sets a framework for a series of subsequent studies in this direction; some of them are presented in the current issue.

The authors of the article *Linguistic profiling of educational and artistic texts* Konstantin V. Voronin, Farida H. Ismaeva, Andrew V. Danilov, present a detailed profiling of adventure stories as fictional texts and contrast them with the texts of educational biographies used in textbooks on Russian as a foreign language. The discriminant parameters which differentiate biographies from textbooks on Russian as a foreign language and adventure stories are as follows: global and local repetitions of nouns and personal pronouns, distribution of nouns in prepositional and genitive cases, past and present verbs. The genre specificity of biography is wider reference ranges of prepositional and genitive cases of nouns and greater connectivity. The research is carried out on a very representative material, includes a detailed analysis of 15 linguistic parameters calculated with the help of RuLingva and a consistent description of the research methodology. The article is an example of cross-genre profiling.

The article *Lexical enrichment of philology textbooks: corpus and statistical approaches* by Khalida N. Galimova, Ekaterina V. Martynova, Svetlana A. Moskvitcheva analyzes the lexical content of Russian language and literature textbooks. As other articles of this block, the article considers textbooks for grades 5–7 of the Russian secondary school, 66 textbooks with a total volume of more than 1.5 million words. The corpus is representative as it contains all textbooks included in the Federal State Educational Standard.

The authors study the vocabulary of textbooks in terms of volume, frequency, and dynamics from grade to grade. One of the noteworthy results is that the

largest vocabulary composition is in the 5th grade textbooks. It seems that this data should still be conceptualized in the light of the general concept of secondary education in Russia.

The authors describe the frequency dictionaries for each grade. The problem of analyzing rare words is also discussed. The obtained frequency dictionaries and "enrichment" dictionaries are divided by thematic groups. The dynamics of the vocabulary composition of textbooks is of particular interest. It turned out that in the 6th grade textbooks, compared to those for the 5th grade, 25% of new words are obsolete (historicisms and archaisms), which preserves Russian cultural code. The authors conclude that the vocabulary of the subjects "Russian Language" and "Literature" is an important material for educating a modern person and preserving cultural traditions of Russia.

The article *Theory of Russian orthography in educational literature for students of the Republic of Belarus* by Evgeniy E. Ivanov, Vladimir I. Kulikovich characterizes teaching Russian orthography in Belarusian universities. The specifics of different textbooks on orthography is whether they introduce it with the help of examples and simple rules or form its theoretical foundations and a fundamental methodological base. The authors distinguish three groups of textbooks based on specific representation of orthography as a theoretical discipline and conclude that students who study orthography within the theoretical approach make fewer errors in writing than those who used textbooks with examples only. At the same time, both groups of students are successful in doing tests.

The article also investigates the issue of unity or discrepancies in the definitions of orthographic concepts in different textbooks. An illustrative example is given when one textbook attributes one meaning to the term 'orthography' and another textbook attributes four (!). The authors assess this situation as follows: "the terminological basis of Russian orthography in Belarusian textbooks <...> in many cases is unscientific". As a result of this research, the authors propose to present the modern theory of Russian orthography considering four basic principles: *systematicity, anthropologism, semantic integrity, and expediency*.

In general, the authors propose their approach to teaching Russian abroad, especially in countries with a large proportion of Russian-speaking population. The approach considers the variability of orthography and other branches of linguistics. This issue seems to be insufficiently studied. The ideas of the article can also be applied to teaching Russian as a foreign language in Russia.

The article *Language of Russian textbooks: diachronic linguistic profiling* by Roman V. Kupriyanov, Gulnoza N. Shoeva, Oksana I. Aleksandrova presents systematic quantitative comparison of texts in Russian language textbooks for grade 5 used in the USSR and Russia in 1937–2015. 24 linguistic parameters of the texts were studied. The RuLingva profiler was used for quantitative analysis, and it revealed interesting patterns of change in educational texts over time. In particular, the authors found unexpectedly that the texts of modern textbooks are simpler (they use shorter sentences and words). Other parameters demonstrate sta-

tistically significant differences between Soviet and Russian textbooks. The article also draws attention to the fact that textbooks have semantic fragments that differ in their linguistic parameters: presentation of theoretical material, exercises, and tasks. The meanings of 24 linguistic parameters in these fragments are analyzed. The article points out that future research can increase the number of textbooks under consideration both by classes and by subjects. This article is a sample of research in this area.

The article *Methods of Anglicisms Monitoring in Discourse of the Russian Youth* by Irina V. Privalova, Anna A. Petrova, Luiza N. Gishkaeva presents the authors' methods of researching anglicisms in contemporary Russian youth discourse. The authors give the results of three surveys conducted at Saratov, Volgograd, and Kazan Universities over the last 7 years. Several hundred respondents took part in the surveys, and the frequency of more than 1300 words of youth sociolect was studied. The authors came to the following conclusions. Anglicisms are significantly superior to other types of words in the youth sociolect. They are primarily rooted in communication among friends and family members and in Internet communication. The frequency of the lexemes in the Russian National corpus is lower than that in real use in the youth environment. This is because the Russian National Corpus lags the real usage; it takes time to fix new units in the language. The paper also shows the influence of foreign and Russian TV series on youth slang.

The new and dynamic phenomenon youth slang is a complex research issue which requires constant monitoring of the situation. This paper is one of the few systematic studies in this area. At the same time, the research has several limitations. Firstly, only Russian national corpus was used; in the future, it is necessary to cross-check the results obtained on other corpora, e.g. Google Books Ngram. Secondly, the book *Dictionary of youth slang* by Shamne & Rebrina[13] was used as a source of words of youth sociolect. The authors note that, "The method of solid sampling in alphabetical order was the most effective in terms of selecting lexemes". It seems that it is necessary to expand the studied vocabulary and to revise the dictionary by Shamne & Rebrina. Some words such as flash drive, fan, content, are beyond just youth usage, they have long been fixed in the language. Consequently, it is difficult to distinguish between youth slang and words of literary language. Finally, a special problem is the problem of homonymy. For example, the word *to bomb* has different meanings in youth and media discourse. The importance of this direction is determined by the urgent task of preserving the Russian language.

Conclusion

Modern linguistics is successfully turning to interdisciplinary approaches to solve the problems it is facing. Linguistic profiling tools based on the achieve-

[13] Shamne, N.L., & Rebrina, L.N. (2017). *Dictionary of youth slang*. Volgograd: Volgy publ.

ments of linguistic statistics, computational linguistics, and artificial intelligence are becoming increasingly relevant. The methodological basis of formalized methods of text analysis is provided by the discoveries made in the field of text theory, functional stylistics, stylistic statistics, and computational linguistics.

References

Blinova, O., & Tarasov, N. (2022). A hybrid model of complexity estimation: Evidence from Russian legal texts. *Frontiers in Artificial Intelligence*, *5*. http://doi.org/10.3389/frai.2022.1008530

Chang, T.A., Arnett, C., Tu, Z., & Bergen, B.K. (2023). When is multilinguality a curse? language modeling for 250 high-and low-resource languages. *arXiv preprint.* https://doi.org/10.48550/arXiv.2311.09205

Corlatescu, D., Ruseti S., & Dascalu, M. (2022). ReaderBench: Multilevel analysis of Russian text characteristics. *Russian Journal of Linguistics, 26*(2), 342–370. https://doi.org/10.22363/2687-0088-30145

Cvrček, V., & Chlumská, L. (2015). Simplification in translated Czech: a new approach to type-token ratio. *Russian Linguistics, 39*, 309–325. https://doi.org/10.1007/s11185-015-9151-8

Dmitrieva, A., Laposhina, A., & Lebedeva, M. (2021). A comparative study of educational texts for native, foreign, and bilingual young speakers of russian: are simplified texts equally simple? *Frontiers in Psychology, 12*, 703690. https://doi.org/10.3389/fpsyg.2021.703690

Gatiyatullina, G., Solnyshkina, M., Solovyev, V., Danilov, A., Martynova, E., & Yarmakeev, I. (2020). Computing Russian morphological distribution patterns using RusAC online server. In *2020 13th International Conference on Developments in eSystems Engineering (DeSE)* (pp. 393–398). IEEE Publ. https://doi.org/10.1109/DeSE51703.2020.9450753

Golovin, B.N. (1971). *Language and statistics*. Moscow: Prosveshchenie Publ. (In Russ.).

Karakanta, A., Dehdari, J., & van Genabith, J. Neural machine translation for low-resource languages without parallel corpora. *Machine Translation, 32*, 167–189. https://doi.org/10.1007/s10590-017-9203-5

Kolmogorova, A.V., Kolmogorova, P.A., & Kulikova, E.R. (2024). About the past, but at different times: computer analysis of textbooks on the history of the USSR / Russia for six generations of students. *Tomsk State University Journal of Philology*, (89), 73–103. (In Russ.). http://doi.org/10.17223/19986645/89/4

Kormilitsyna, M.A., & Sirotinina, O.B. (2013). Functional stylistics and its place in modern linguistics. In L.R. Duskaeva (Ed.), *Slavic stylistics. The 21st century: collection of articles* (pp. 101–111). Saint Petersburg: SPbU Publ. (In Russ.).

Kozhina, M.N. (1989). On functional semantic-stylistic categories in the aspect of the communicative theory of language. In *Varieties and genres of scientific prose. Linguostylistic features* (pp. 3–27). Moscow: Nauka Publ. (In Russ.).

Krongauz, M.A. (2009). *Russian language on the verge of a nervous breakdown*. Moscow: Languages of Slavic cultures Publ. (In Russ.).

Kupriyanov, R.V., Solnyshkina, M.I., Dascalu, M., & Soldatkina, T.A. (2022). Lexical and syntactic features of academic Russian texts: a discriminant analysis. *Research Result.*

Theoretical and Applied Linguistics, 8(4), 105–122. http://dx.doi.org/10.18413/2313-8912-2022-8-4-0-8

Kuznetsova, I. (2015). *Linguistic profiles: going from form to meaning via statistics.* De Gruyter Mouton. http://doi.org/10.1515/9783110361858

Laposhina, A.N., Veselovskaya, T.S., Lebedeva, M.Yu., & Kupreshchenko, O.F. Lexical composition of the Russian language textbooks for primary school: corpus study. In *Computational linguistics and intellectual technologies: based on the materials of the international conference "Dialogue 2019". Vol. 18* (pp. 351–363). (In Russ.).

Laposhina, A.N., & Lebedeva, M.Yu. (2021). Textometer: an online tool for determining the difficulty level of a text in Russian as a foreign language. *Russian Language Studies, 19*(3), 331–345. (In Russ.). http://doi.org/10.22363/2618-8163-2021-19-3-331-345

Lipmann, W. (1922). *Public Opinion.* New York: Macmillan.

Lukashevich, N.V., & Dobrov, B.V. (2015). Designing linguistic ontologies for information systems in broad subject areas. *Ontology of Designing,* (1), 47–69.

Lyashevskaya, O.N., & Sharov, S.A. (2009). *Frequency Dictionary of the Modern Russian Language (based on materials from the Russian National Corpus).* Moscow: Azbukovnik Publ. (In Russ.).

Lyashevskaya, O., Panteleeva, I., & Vinogradova, O. (2021). Automated assessment of learner text complexity. *Assessing Writing, 49,* 100529. https://doi.org/10.1016/j.asw.2021.100529

McNamara, D.S., Graesser, A.C., McCarthy, P.M., & Cai, Z. (2014). Automated Evaluation of Text and Discourse with Coh-Metrix. Cambridge University Press.

Mikheev, M.Yu., & Erlich, L.I. (2018). Idiostyle profile and determination of text authorship by frequencies of function words. *Automatic Documentation and Mathematical Linguistics,* (2), 25–34. (In Russ.).

Morozov, D.A., Glazkova, A.V., & Iomdin, B.L. (2022). Text complexity and linguistic features: Their correlation in English and Russian. *Russian Journal of Linguistics, 26*(2), 426–448. https://doi.org/10.22363/2687-0088-30132

Namestnikov, A.M., Pirogova, N.D., & Filippov, A.A. (2021). An approach to the automatic construction of a linguistic ontology for determining the interests of social network users. *Ontology of design, 11*(3), 351–363. (In Russ.). http://doi.org/10.18287/2223-9537-2021-11-3-351-36

Oborneva, I.V. (2006). *Automated assessment of the complexity of educational texts based on statistical parameters.* (Candidate dissertation, Moscow). (In Russ.).

Paraschiv, A., Dascalu, M., & Solnyshkina, M.I. (2023). Classification of Russian textbooks by grade level and topic using ReaderBench. *Research Result. Theoretical and Applied Linguistics, 9*(1), 50–63. https://doi.org/10.18413/2313-8912-2023-9-1-0-4

Sakhovskiy, A., Solovyev, V., & Solnyshkina, M. Topic modeling for assessment of text complexity in Russian textbooks. In *Proceedings of 2020 Ivannikov Ispras Open Conference (ISPRAS)* (pp. 102–108). IEEE Publ. https://doi.org/10.1109/ISPRAS51486.2020.00022

Saussure, F. de. (1977). Trudy po iazykoznaniiu [Writings in General Linguistics]. Moscow: Progress, 695 p.

Serdobolskaya, N.V., & Toldova, S.Yu. Evaluation predicates: type of evaluation and syntax of the construction. In *"Computer linguistics and intellectual technologies": proceedings of the International Conference 'Dialogue' 2005* (pp. 436–443). Moscow: Nauka Publ. (In Russ.).

Solnyshkina, M.I., Solovyev, V.D., Gafiyatova, E.V., & Martynova, E.V. (2022). Text complexity as an interdisciplinary problem. *Issues of Cognitive Linguistics,* (1), 18–39. https://doi.org/10.20916/1812-3228-2022-1-18-39

Solovyev, V., Ivanov, V., & Solnyshkina, M. (2018). Assessment of reading difficulty levels in Russian academic texts: Approaches and metrics. *Journal of Intelligent & Fuzzy Systems, 34*(5), 3049–3058 http://doi.org/10.3233/JIFS-169489

Solovyev, V., Solnyshkina, M., & McNamara, D. (2022). Computational linguistics and discourse complexology: Paradigms and research methods. *Russian Journal of Linguistics, 26*(2), 275–316. https://doi.org/10.22363/2687-0088-31326

Toldova, S., Anastasiya, A.B., Lyashevskaya, O., & Ionov, M. (2015). Evaluation for morphologically rich language: Russian NLP. In *Int'l Conf. Artificial Intelligence. ICAI'15* (pp. 300–306).

Valeev, A., Gibadullin, I., Khusainova, A., & Khan, A. (2019). Application of Low-resource Machine Translation Techniques to Russian-Tatar Language Pair. *arXiv preprint.* http://doi.org/10.48550/arXiv.1910.00368

Vinogradov, V.V. (1938). Modern Russian language. Grammatical doctrine of the word. Moscow; Leningrad State educational-pedagogical publishing house of the People's Commissariat of Education of the RSFSR. (In Russ.).

Virk, S.M., Hammarström, H., Borin, L., Forsberg, M., & Wichmann, S. (2020). From Linguistic Descriptions to Language Profiles. In *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)* (p. 23–27). Marseille: European Language Resources Association Publ.

Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent Trends In Deep Learning Based Natural Language Processing. *IEEE Computational intelligence magazine, 13*(3), 55–75. http://doi.org/10.1109/MCI.2018.2840738

Zinder, L.R., & Stroeva, T.V. (1968). Historical morphology of the German language. Leningrad: Prosveshchenie Publ. (In Russ.).

Bio notes:

Marina I. Solnyshkina, Doctor Habil. of Philology, Professor of the Department of Theory and Practice of Teaching Foreign Languages, Head of "Multidisciplinary Text Investigation" Research Lab, Institute of Philology and Intercultural Communication, Kazan Federal University, 18 Kremlevskaya St, Kazan, 420008, Russian Federation. *Research interests:* text analytics, text complexity, discursive complexology, lexicography, linguistic personality. ORCID: 0000-0003-1885-3039. SPIN-code: 6480-1830. Researcher ID: E-3863-2015. Scopus ID: 56429529500. E-mail: mesoln@yandex.ru

Valery D. Solovyev, Doctor Habil. of Physical and Mathematical Sciences, Professor, a member of Presidium of Multidisciplinary Association for Cognitive Research, the author of four monographs and over 70 publications on text complexity, Chief Researcher of "Multidisciplinary Text Investigation" Research Lab, Institute of Philology and Intercultural Communication, Kazan Federal University, 18 Kremlevskaya St, Kazan, 420008, Russian Federation. *Research interests:* cognitive science, computational linguistics, artificial intelligence. ORCID: 0000-0003-4692-2564. SPIN-code: 5791-3820. Researcher ID: C-8023-2015. Scopus ID: 26665013000. E-mail: maki.solovyev@mail.ru

Yulia N. Ebzeeva, Doctor of Social Sciences, PhD in Philology, First Vice-Rector — Vice Rector for Education and Head of Foreign Language Department, RUDN University, 6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation. *Research interests*: French lexicology and stylistics, translation studies, intercultural communication, sociolinguistics, migration studies and educational policy. ORCID: 0000-0002-0043-7590. SPIN-code: 3316-4356. E-mail: ebzeeva-jn@rudn.ru

DOI: 10.22363/2618-8163-2024-22-4-501-517
EDN: AMYSNF

Вступительная статья

Подходы и инструменты лингвистического профилирования текста на русском языке

М.И. Солнышкина[1], В.Д. Соловьев[1], Ю.Н. Эбзеева[2]

[1]Казанский (Приволжский) федеральный университет, *Казань, Российская Федерация*
[2]Российский университет дружбы народов, *Москва, Российская Федерация*

✉ mesoln@yandex.ru

Аннотация. Развитие подходов и усовершенствование инструментов оценки лингвистической и когнитивной сложности учебного текста востребовано как в науке, так и практике обучения. Особую значимость прогнозирование трудностей восприятия и понимания, а также ранжирование текстов по классам, т.е. количеству лет формального обучения, или уровням владения языком (A1–C2) имеет в системе образования. Цель исследования — продемонстрировать, каким образом современные методологии, алгоритмы и инструменты аналитики текстов на русском языке реализованы в автоматическом анализаторе RuLingva, а также представить статьи тематического выпуска, посвященного комплексному анализу учебников по русскому языку для российских и белорусских школ. Показано, что современная парадигма дискурсивной комплексологии опирается на разработанные в российском языкознании методы стилостатистики, позволяющие выявлять функциональные характеристики языковых единиц и осуществлять их верификацию на материале больших языковых данных. Функционирующие на портале RuLingva сервисы предназначены для преподавателей и исследователей и позволяют в автоматическом режиме не только осуществлять аналитику учебного текста, но и прогнозировать его целевую аудиторию на основании данных о читабельности, лексическом разнообразии, абстрактности, частотности, терминологической плотности. В режиме «Русский как иностранный» RuLingva выгружает из текста списки слов, соответствующие каждому из уровней владения языком, и оценивает долю каждого из них, предоставляя таким образом материал для пред- и посттекстовой работы преподавателя. Алгоритм функционирования RuLingva разработан на основе типологии учебных текстов и имеет в качестве перспективы создание функционала оценки вербального интеллекта и читательской грамотности обучающегося. Перспектива развития RuLingva связана с расширением спектра предикторов сложности и внедрением функции автоматического определения предметной области учебного текста. Оба направления планируется реализовать при помощи нейронных сетей и созданных на их основе классификационных моделей, а также на базе «типологических паспортов» учебных текстов различной сложности и тематической направленности.

Ключевые слова: лингвистический анализ, текстовый профайлер, RuLingva, сложность текста, учебный текст, типологический паспорт текста, предикторы сложности

Вклад авторов: Солнышкина М.И. — разработка концепции, проведение исследования, подготовка и редактирование текста; Соловьев В.Д. — разработка методологии, проведение исследования; Эбзеева Ю.Н. — проведение исследования, утверждение окончательного варианта статьи.

Финансирование. Работа выполнена за счет средств Программы стратегического академического лидерства Казанского (Приволжского) федерального университета (ПРИОРИТЕТ–2030). Работа выполнена в рамках проекта № 050738-0-000 системы грантовой поддержки научных проектов РУДН.

Конфликт интересов: Авторы заявляют об отсутствии конфликта интересов.

История статьи: поступила в редакцию 02.07.2024; принята к печати 18.08.2024.

Для цитирования: *Солнышкина М.И., Соловьев В.Д., Эбзеева Ю.Н.* Подходы и инструменты лингвистического профилирования текста на русском языке // Русистика. 2024. Т. 22. № 4. С. 501–517. http://doi.org/10.22363/2618-8163-2024-22-4-501-517