DOI: 10.18413/2313-8912-2022-8-4-0-8

Roman V. Kupriyanov[1]
Marina I. Solnyshkina[2]
Mihai Dascalu[3]
Tatyana A. Soldatkina[4]

Lexical and syntactic features of academic Russian texts: a discriminant analysis

[1] Text Analytics Laboratory, Kazan Federal University
18 Kremlevskaya St., Kazan, 420008, Russia
Kazan National Research Technological University
68 Karl Marx St., Kazan, 420015, Russia
E-mail: kroman1@mail.ru

[2] Text Analytics Laboratory, Kazan Federal University,
18 Kremlevskaya St., Kazan, 420008, Russia
E-mail: mesoln@yandex.ru

[3] Polytechnic University of Bucharest
313 Splaiul Independentei St., Sector 6, Bucharest, 060042, Romania
E-mail: mihai.dascalu@upb.ro

[4] Research Lab Laboratory "Expert Systems for Processing Language Structures and Vibroacoustics", Kazan Federal University
18 Kremlevskaya St., Kazan, 420008, Russia
Mari State University
44 Kremlevskaya St., Yoshkar-Ola, 424000, Russia
E-mail: fia.vr.solta@gmail.com

Received 29 August 2022; accepted 12 December 2022; published 30 December 2022

Acknowledgements. This paper has been supported by the Kazan Federal University Strategic Academic Leadership Program ("PRIORITY-2030"), Strategic Project №4.

We thank Polina Alexandrovna Lekhnitskaya, a student at Kazan Federal University, for her assistance in compiling the corpus of academic texts and cooperation while conducting the research.

Abstract. This article presents three mathematical models to differentiate academic texts from three subject discourses written in Russian (i.e., Philological, Mathematical, and Natural Sciences) which further enable design and automated profiling of corresponding typologies. Our models include 5 indices, one at surface level (i.e., sentence length) and 4 syntax features (i.e., mean verbs per sentence, mean adjectives per sentence, local noun overlap, and global argument overlap). We identified and validated the five statistically significant features out of 45 linguistic features extracted from our research corpus consisting of 91.185 tokens. The shortest sentence length is found in Russian language textbooks while the longest sentences are identified in Natural Science texts. The mean number of verbs, nouns, and adjectives per sentence is higher in Natural Science textbooks, whereas Mathematics

discourse is characterized by the shortest word length, highest local noun overlap, and highest global argument overlap. We assign the metric differences between the three discourses to their functions: Natural Science texts are characterized by descriptions and narrative passages in contrast to Philology that is associated with opinions. Mathematical discourse operates with precise definitions, explanations and justifications thus exercising numerous overlaps. The discriminant analysis built on top of the features supports the development of text profilers targeting parametric analyses. The automation of these features and the provided formulas for classification enable the design and development of text profilers required for textbook writing and editing. Our findings are useful for professional linguists, technologists, and academic writers to select and modify texts for their target audience.

Keywords: Typology; Lexical features; Automation profilers; Subject domain; Syntactic features; Mathematical model; Discriminant analysis

How to cite: Kupriyanov, R. V., Solnyshkina, M. I., Dascalu, M. and 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. DOI: 10.18413/2313-8912-2022-8-4-0-8

DOI: 10.18413/2313-8912-2022-8-4-0-8

Куприянов Р. В.[1]
Солнышкина М. И.[2]
Даскалу М.[3]
Солдаткина Т. А.[4]

Лексические и синтаксические параметры академического текста: дискриминантный анализ

[1] НИЛ «Текстовая аналитика», Казанский (Приволжский) федеральный университет
ул. Кремлевская, 18, Казань, 420008, Россия
Казанский национальный исследовательский технологический университет (КНИТУ)
ул. Карла Маркса, 68, Казань, 420015, Россия
E-mail: kroman1@mail.ru

[2] НИЛ «Текстовая аналитика», Казанский (Приволжский) федеральный университет
ул. Кремлевская, 18, Казань, 420008, Россия
E-mail: mesoln@yandex.ru

[3] Бухарестский политехнический университет
Splaiul Independentei 313, Sector 6, Bucharest, 060042, Romania
E-mail: mihai.dascalu@cs.pub.ro

[4] НИЛ «Экспертные системы обработки языковых конструкций и виброакустика, Казанский (Приволжский) федеральный университет
ул. Кремлевская, 18, Казань, 420008, Россия
Марийский государственный университет
ул. Кремлевская, 44, Йошкар-Ола, 424000, Россия
E-mail: fia.vr.solta@gmail.com

Статья поступила 29 августа 2022 г.; принята 12 декабря 2022 г.;
опубликована 30 декабря 2022 г.

Информация об источниках финансирования или грантах, благодарности: Работа выполнена за счет средств Программы стратегического академического лидерства Казанского (Приволжского) федерального университета («ПРИОРИТЕТ-2030»), Стратегического проекта №4.

Мы благодарим Лехницкую Полину Александровну, студентку Казанского федерального университета, за помощь в подготовке корпусов учебных текстов и проведении исследования.

Аннотация. В статье представлены математические модели дифференциации академических текстов трех предметных дискурсов на русском языке (филологического, математического и естественнонаучного), которые являются основой разработки и автоматизации профилирования текстов. Наша модель включает индексы двух групп параметров, а именно, поверхностных (например, длина предложения) и синтаксических (например, среднее значение глаголов в предложении, среднее значение прилагательных в предложении, локальный повтор существительных и глобальный повтор аргументов). Мы определили и подтвердили 5 статистически значимых признаков из 45 лингвистических признаков, извлеченных из нашего исследовательского корпуса, состоящего из 91185 токенов. Дискриминантный анализ, осуществленный на основе этих функций, подтвердил валидность профилирования текстов основанного на параметричесом анализе. Наши результаты будут полезны профессиональным лингвистам, программистам и разработчикам учебных и контрольно-измерительных материалов при выборе и модификации текстов для целевой аудитории.

Ключевые слова: Профилирование текста; Лексические признаки; Автоматические профайлеры; Предметная область; Синтаксические признаки; Математическая модель; Дискриминантный анализ

Информация для цитирования: Куприянов Р. В., Солнышкина М. И., Даскалу М., Солдаткина Т. А. Лексические и синтаксические параметры академического текста: дискриминантный анализ // Научный результат. Вопросы теоретической и прикладной лингвистики. Т. 8. № 4. С. 105-122. DOI: 10.18413/2313-8912-2022-8-4-0-8

Introduction

The modern paradigm of applied linguistics addresses numerous problems – for example, translation algorithms, Natural Language Processing, and text mining (lit.Russ. "intelligent" text analysis). Results of applied linguistics research are extensively applied, from selecting texts with the designated content to recommendations for modifying the text for a certain category of potential readers. At present, researchers and users have readily available several automated text analyzers like TextInspector, Lextutor, Coh-Metrix, ReaderBench, Textometer, and RuLingva. These systems compute more than 200 text features and provide researchers with materials for describing, comparing, and altering texts depending on the users' linguistic-pragmatic goals. For example, LexTutor (https://www.lextutor.ca/vp/eng/) classifies vocabulary by origin, while TextInspector (https://textinspector.com/) considers the Common European scale (eng. CEFR) (https://www.coe.int/ru/web/lang-migrants/cefr-and-profiles). However, none of the existing analyzers is a discourse profiler – i.e., they do not define the register, discourse, and the type of a text based on its linguistic

features. Of interest for these systems is the identification of reference ranges for the parameter metrics which enables the classification of text types. The demand for such profilers is especially high when selecting texts for specific purposes (e.g., educational, monitoring, informative, suggestive), as well as for authorship identification or the selection of text materials for various categories of users.

This article aims to establish a list of typological linguistic features that differentiate texts corresponding to three subject discourses (i.e., Mathematical, Philological, and Natural Science). The mathematical model relies on a discriminant analysis that enables follow-up automated text profiling (i.e., the attribution of a text to a certain level of complexity and discourse type).

The **hypothesis** of this study is that academic texts of a predefined complexity (i.e., within the range of one academic year) and intended for use in various subject areas (i.e., Mathematics, Philology, and Natural Science) exhibit quantitative lexical and syntactic differences. Such differences are typological in nature and can lead to the identification of discourse type and even the author (source). The identification of discriminative features and the design of a mathematical model of the text further facilitate the inter- and intra-discourse classification of texts.

Our main research objective is to develop a mathematical model to predict the complexity of academic texts in Russian using a limited list of linguistic features. We also aim at providing researchers with a practical and reproducible route to developing new language resources for Russian as a low-resource language.

Literature review

The specificity of an academic (or educational or scientific) text lies in its communicative function and pragmatic component, namely, in its focus on comprehension from the target audience. Zherebtsova (2007: 29) emphasizes the importance of information transfer when defining an educational and scientific text as a written message characterized by semantic and structural completeness. As such, the information content of a text as a unit of discourse is largely determined by its linguistic features: morphological, lexical, syntactic, and discourse (Solnyshkina, Harkova, Kazachkova, 2020). These features reveal the specifics of the educational and scientific text in different ways, determining its perceived difficulty for various categories of linguistic profiles (Solnyshkina, Kazachkova, Harkova, 2020).

The perception of difficulty for oral and printed (electronic) texts is correlated to quantitative features that include text length, syllable or character means for words (i.e., word length), or word counts in a sentence (i.e., sentence length). These features are considered in statistical analyses of text complexity and their correlation to text difficult is linked to the capacity of working memory (Oborneva, 2006: 5).

Sentence length as a predictor of complexity is of particular interest because it relates to syntax. Inherently, syntax may be more complex for sentences with an increased number of words; thus, high values for this feature are indicative of potential difficulties in understanding the text (McNamara, Graesser, McCarthy & Cai, 2014: 2). Word length is evaluated in a similar manner: longer words require more time to comprehend, work with, and store for a short term (Vakhrusheva, Solnyshkina, Kupriyanov, Gafiyatova, Klimagina, 2021: 15). Shorter words are easier to read; moreover, they are easier to comprehend and disambiguate since they tend to have fewer senses (Kiselnikov, 2015: 4).

Other morphological features also play an important role in text comprehension – for example, the proportions of various parts of speech in the text. Corpus linguistics has developed methods for identifying genres based on the relative frequencies of individual parts of speech (Seifart, Danielsen, Meyer, Nordhoff et al., 2012: 10). Statistically

significant differences in registers and types of discourses were validated in several languages (Biber, 2006: 261). For example, verbs overlap was confirmed to create a more cohesive event structure that is easier to comprehend using the situational model; this parameter is especially relevant in the analysis of narrative texts (McNamara, Graesser & Louwerse, 2012: 89–116). Similar patterns have also been identified in Russian texts (Zhuravlev, 1988: 84–150; Sirotinina, 2009: 312). For example, the mean adjective and noun counts, as well as the genitive case, were validated as reliable complexity predictors. The increase of genitive cases in biology texts from the 5th to the 11th grade is 7% (from 34% to 41%), while social science texts exhibit a more drastic increase from 23% to 38% (Gatiyatullina, Solnyshkina, Solovyev, Danilov et al., 2020: 393–398).

Linguistic features of text complexity also include relative predictors (i.e., measures based on the relation of specific groups of units to others): the nominative ratio of verbs to nouns and the descriptive ratio of adjectives to nouns (Martynova, Solnyshkina, Merzlyakova, Gizatulina, 2020: 72–80).

Lexical features relate to the overlap of individual lexemes. Research indicates the significance of local overlap of nouns – i.e., repetitions of the same lexeme within one sentence or in adjacent sentences -, as well as global repetitions within the entire text (Corlatescu, Ruseti & Dascalu, 2022: 354; McNamara, Graesser, McCarthy & Cai, 2014: 2). Similar features include local and global argument overlaps (Crossley, Varner, Roscoe & McNamara, 2013: 3) that consider noun, pronoun, or a noun phrase in one sentence as a co-referent of a noun, pronoun, or a noun phrase in another sentence (McNamara, Graesser, McCarthy & Cai, 2014: 90).

Researchers also highlight lexical diversity (TTR - Type Token Ratio; the ratio of words to word forms) as a complexity predictor (Graesser, McNamara, Louwerse & Cai, 2004: 194). With TTR=1.0, none of the words in the text are repeated; however, such texts are not natural since the absence of lexical repetitions increases the difficulty of texts. Low TTR values (< 0.5) indicate a high repetition of words, which positively impacts text processing. The target audience of these texts consists of speakers with a limited vocabulary, namely foreign language learners or young students (Malvern, Richards, Chipere & Durán, 2004). TTR is measured on texts no longer than 1000 tokens as lengthier sequences result in an increase of functional words on one hand, and a decrease in content words, on the other. TTR values measured on texts longer than 1000 tokens are considered unreliable; thus, these texts need to be divided into fragments on which TTR is measured separately (Vakhrusheva, Solnyshkina, Kupriyanov, Gafiyatova, Klimagina, 2021: 88–99).

A validated predictor for the complexity of academic texts is the Flesch-Kincaid Readability Index (FK), originally developed for texts in English (Flesch, 1948: 221–233) and adapted for the Russian language only at the beginning of this century (Solnyshkina and Kiselnikov, 2015). The popularity of this index was facilitated by two factors: ease of calculations (and subsequent successful automation) and its match to the academic age of the reader (i.e., the number of years of formal schooling). Currently, this formula is successfully used for a variety of purposes from matching books to reader vocabularies, to predicting the success of a website. This text readability index is measured based on two basic metrics – mean sentence length and mean word length (Solnyshkina and Kiselnikov, 2015). FK is widely used to assess text appropriateness for different categories of readers pertaining to the military, medical institutions, insurance companies, and even car dealerships.

The two most notable readability formulas for the Russian language were designed for texts containing various discourse types. First, FC (SIS) (eq. 1) was developed on the corpus of academic texts and validated in psycholinguistic experiments with school children:

(1) FC(SIS) = 208.7 – 2.6 × ASL – 39.2 × ASW

where ASL is the mean sentence length in tokens and ALS is the mean word length in syllables (Solovyev, Ivanov & Solnyshkina, 2018).

Second, the readability formula of Oborneva (2006) [FC(O), eq. 2] was developed on fiction texts; however, it provides overestimated results when applied to texts of other types:

(2) FC(O) = 206.835 – (1.3× ASL) – (60.1× ASW) (see Solnyshkina, McNamara & Zamaletdinov, 2022).

The index of abstractness is also recognized by many researchers as a complexity predictor (Solovyev, Ivanov & Akhtiamov, 2019: 215–227) since abstract concepts hinder text comprehension (Solnyshkina and Kiselnikov, 2015). This parameter is especially significant in the complexity assessment of texts intended for younger students who more easily understand concrete words and may struggle with abstract concepts (Vakhrusheva, Solnyshkina, Kupriyanov, Gafiyatova, Klimagina, 2021: 15).

Methods

The set of features as listed above enables not only to carry out a multi-factor analysis of the linguistic complexity of the text but also to define a profile of the text using a limited number of features (i.e., assign it to a certain type, discourse, and level of complexity).

The starting point for this study was the idea that academic texts exhibit a quantitative typology, namely their "homogeneity" to teach a certain subject to students of a certain grade. Typology as a method is based on the concept of “fuzzy sets” of elements in which the transition of an element (in our case, a text) from one class (category) to another is carried out gradually. Elements of a class possess two types of features: inherent features (i.e., features typical of a class) and specific, individual features. The transition of an element from one class to another implies the accumulation of typological features of another set. For example, the complexity of Mathematics texts for the 2nd and the 3d grades is supposed to be different, although linguistic differences between them are few and minor. However, these differences may be elicited in the metrics of morphological, lexical, and syntactic features. In contrast, when considering texts of the same complexity but different subject areas (e.g., texts used to teach Russian in the 2nd grade and texts used to teach Mathematics in the 2nd grade) we assume they differ in several features. Moreover, the list of these features may differ when comparing texts of the same subject, but of varying complexity.

Our study was carried out in three stages described subsequently:

(1) Preparation, cleaning, and corpus pre-processing

The corpus for this study was compiled from seven textbooks on three subjects (the Russian language, Mathematics, and Science) from the Federal list of textbooks of the Russian Federation (https://fpu.edu.ru/, Order of the Ministry of Education of Russia No. 254, May 20, 2020), summing up to a total size of 95377 tokens. The selection of the books was performed based on the expert opinion of teachers practicing in primary schools. The sub-corpora for the 3 subjects were balanced in terms of their size (see Table 1).

Meta-descriptions, prefaces, author’s introductory words, contents, illustrations, inscriptions, phrases like “Figure 1”, notes, self-control questions, laboratory tasks, chapter titles, subheadings, footers, and running headlines were deleted to ensure consistency of the language material at the pre-processing stage. The textbooks were divided into 87 texts of about 750 – 1000 tokens: 20 Mathematics texts, 30 texts from the textbooks used to teach the Russian language, and 37 texts from textbooks on Natural Science. The variation in the sizes of texts under study, i.e. the range of 750 – 1000, was caused by the following: 1) we followed the textbooks segmentation into chapters and

did not add a fragment to a chapter which in Natural science varies within 700 – 1000 words; 2) we did not increase the recommended sample text size to be within the range of 700 – 1000 words (Biber, 2006). We also randomly selected 10 texts (3 in Mathematics, 3 in the Philology, and 4 in Natural Sciences) to test our model. These 10 texts were not used in the discriminant analysis.

Table 1. Corpus Size
Таблица 1. Размер корпуса исследования

Discourse domain	Textbook size (in tokens)	Subcorpus size (in tokens)
Philology	13702	38478
	20384	
	4392	
Maths	16991	28728
	11737	
Science	19770	28171
	8401	

(2) Measurement of metrics of 45 linguistic features with the help of the automatic analyzer RuLingva (https://rulingva.kpfu.ru/) and the analysis of 14 statistically significant features

The metrics of the linguistic features of the texts under study were calculated using the automatic analysis RuLingva (https://rulingva.kpfu.ru/). After the initial screening, we selected 14 of the 45 features calculated with RuLingva in accordance to the previous work performed by Solnyshkina, Solovyev, Gafiyatova, Martynova (2022): sentence length (mean words in a sentence), word length (mean syllables in a word, mean nouns per sentence, the mean verbs per sentence, mean adjectives per sentence, Flesch-Kincaid index – FK(SIS)), index of abstractness, local noun overlap, global noun overlap, local argument overlap, global argument overlap, lexical diversity (Type token ratio, TTR), the nominative ratio of verbs to nouns, the descriptive ratio of adjectives to nouns, number of one-syllable words, number of two-syllable words, number of three-syllable words, number of four-syllable words. All other features (e.g., number of nouns in different cases, number of tense forms of the verb, Flesch-Kincaid index (FK(O)) were excluded from the analysis based on the similarity of the values for these features across all three sub-corpora.

(3) Development of the profiling method based on a discriminant analysis of the metrics of linguistic features.

The statistical analysis of the 14 features from the 87 texts was carried out using STATISTICA. After checking for the normality of the distributions, non-parametric Kruskal-Wallis *H* tests were conducted to assess differences between the blocks since the features were not normally distributed.

The discriminant analysis was employed to identify typological features of the texts and to calculate the formulas for classifying texts by subject discourse. We used the Discriminant Analysis module in STATISTICA and calculated the values of Wilks lambda (λ) and F-criterion. Wilks lambda value (λ) close to 0 indicates good discrimination (i.e., the contrasted objects have statistically significant differences). The *F* value of a variable in contrasted objects also indicates their statistically significant differences, thus being a measure that has a unique contribution to predicting the classification of an element to a group. Thus, we assess the correctness of the classification of the texts from this study based on the values of λ Wilks and the F-criterion.

Research results

In accordance to our research method, 14 features were analyzed in detail. Columns III - V from Table 2 present the means and standard deviations of all features corresponding to the texts under this study. Column VI denotes statistically significant features with an asterisk * ($p < .05$). Kruskal-Wallis H test confirmed that most of the features of the texts, with the exception of the 'abstractness index' and 'global noun overlap' (lines 9 and 11), are statistically significant (see Table 2).

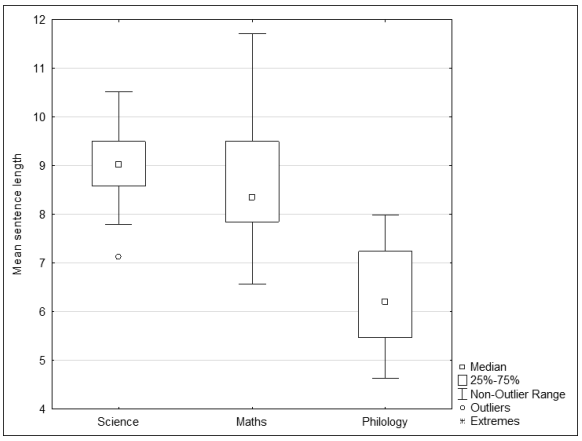
Table 2. Linguistic features of texts of three sub-corpora
Таблица 2. Лингвистические параметры текстов трех предметных подкорпусов

	Parameter	Science ($N = 37$)	Maths ($N = 20$)	Philology ($N = 30$)	Kruskal-Wallis Test $H(2, N = 87)$	p
I	II	III	IV	V	VI	VII
1.	Mean sentence length	9.05±0.73	8.76±1.43	6.30±1.07	53.66	< .01*
2.	Mean word length (in syllables)	2.38±0.18	1.97±0.14	2.27±0.22	51.16	< .01*
3.	Mean of nouns per sentence	3.32±0.36	3.15±0.47	2.55±0.37	40.07	< .01*
4.	Mean verbs per sentence	1.42±0.17	0.97±0.15	0.95±0.19	57.59	< .01*
5.	Mean adjectives per sentence	0.96±0.16	0.74±0.21	0.63±0.15	39.97	< .01*
6.	Nominative ratio	0.43±0.07	0.31±0.03	0.37±0.06	38.54	< .01*
7.	Descriptive ratio	0.29±0.05	0.23±0.05	0.25±0.04	19.90	< .01*
8.	FC index (SIS)	4.83±0.59	2.51±0.81	3.18±0.82	56.53	< .01*
9.	Abstract index	2.60±0.13	2.57±0.12	2.58±0.10	0.61	0.73
10.	Local noun overlap	0.15±0.06	0.39±0.10	0.10±0.04	53.13	< .01*
11.	Global Noun overlap	0.04±0.02	0.03±0.01	0.05±0.07	0.05	0.98
12.	Local argument overlap	0.45±0.13	0.69±0.12	0.28±0.10	54.57	< .01*
13.	Global argument overlap	0.14±0.05	0.08±0.02	0.11±0.07	21.28	< .01*
14.	TTR	0.64±0.05	0.45±0.06	0.60±0.04	50.83	< .01*

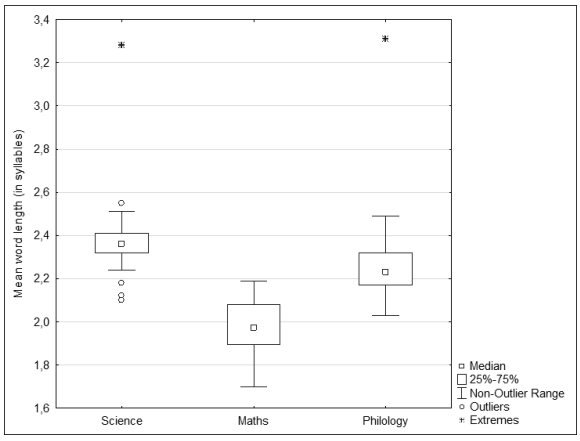
* $p < .05$ — statistically significant differences

We consider for in-depth analysis all features that exhibit statistically significant differences between the three sub-corpora. Based on the data (see Table 2) and range diagrams (see Figures 1 a and b), we argue that the mean sentence length and mean syllables discriminate texts of different subject areas: sentences in Philological texts (sub-corpus of texts used to teach Russian) are the shortest – 6.30±1.07 words, and the longest sentences are in texts in the natural science sub-corpus – 9.05±0.73 words. The shortest words are used by the authors of the Mathematical texts – 1.97±0.14 words, and the longest appear in Natural Science texts – 2.38±0.18 words.

Figure 1. a) Mean sentence length (in words); b) Mean word length (in syllables)
Рисунок 1. a) Средняя длина предложения (в словах); b) Средняя длина слов (в слогах)



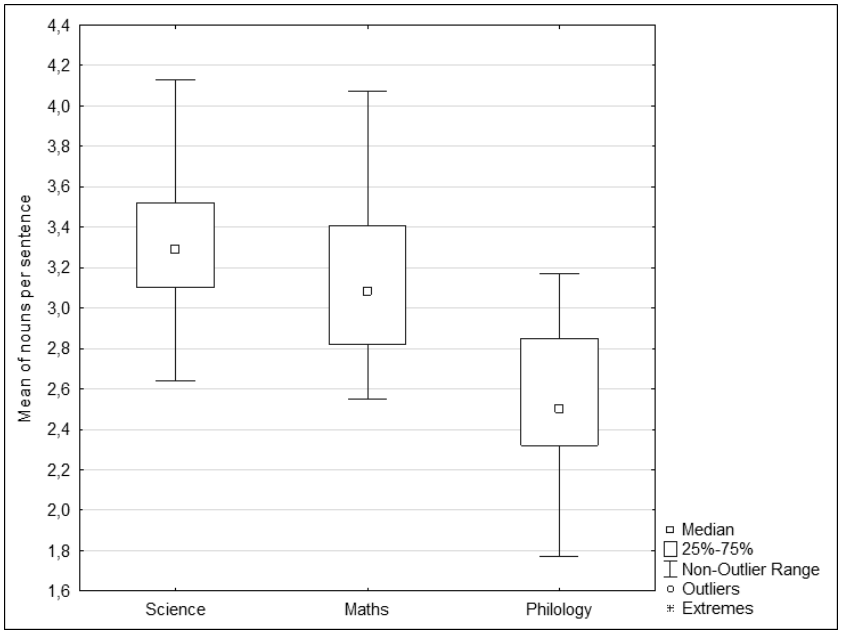
(a)



(b)

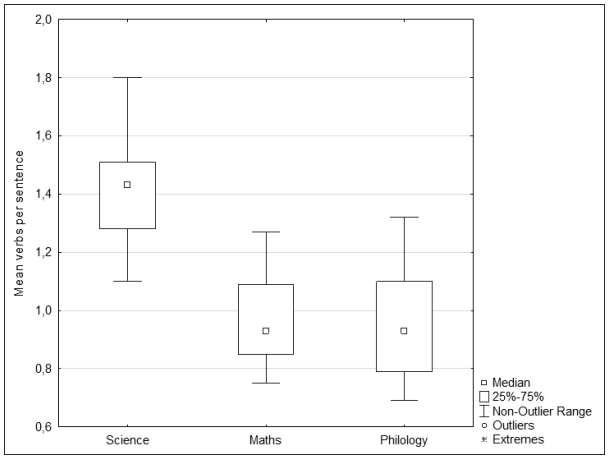
Figure 2 shows that Philological texts differ from the rest of the blocks: the number of nouns per sentence (2.55±0.37) in these texts is the lowest, while the same metric in Natural science (3.32±0.36) and Mathematical (3.15± 0.47) texts differs insignificantly.

Figure 2. Mean nouns per sentence
Рисунок 2. Среднее количество существительных на предложение

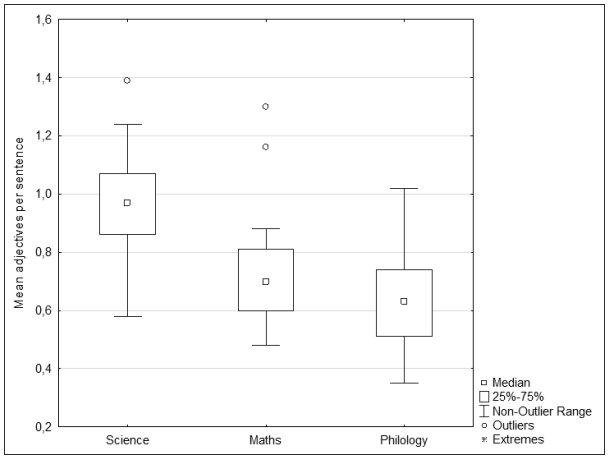


The mean numbers of verbs and adjectives per sentence are also the highest in natural science texts: with verbs having 1.42±0.17 and adjectives 0.96±0.16 per sentence (see Figures 3a and 3b). The differences in these features in texts of Philological and Mathematical sub-corpora are statistically insignificant.

Figure 3. a) Mean verbs per sentence; b) Mean adjectives per sentence
Рисунок 3. a) Среднее количество глаголов на предложение; b) Среднее количество прилагательных на предложение



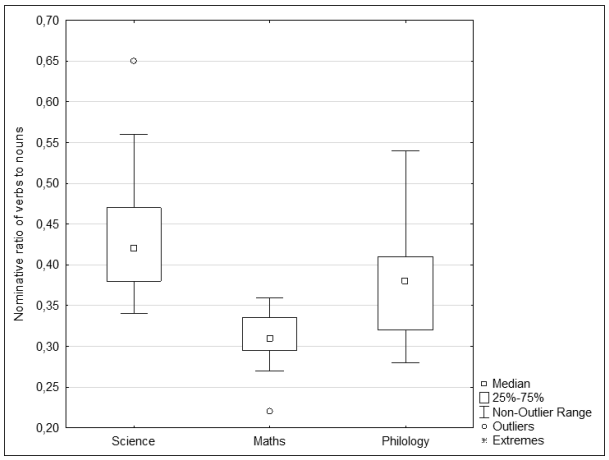
(a)



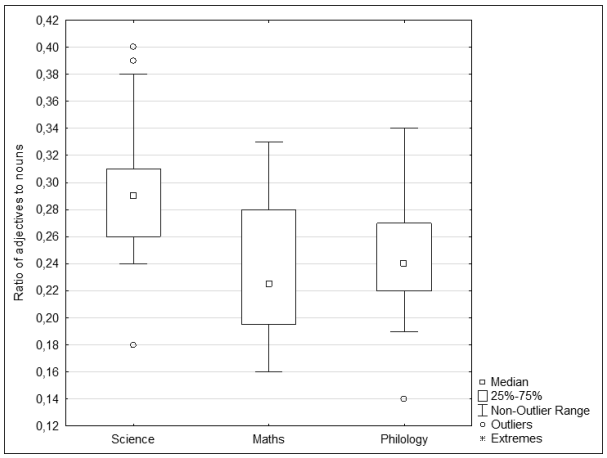
(b)

The nominative ratio of verbs to nouns also exhibits significant differences (see Figure 4 a); the highest values are observed in the natural science corpus (0.43±0.07), while the lowest are in the Mathematical corpus (0.31±0.03). The Philological and Natural Science texts have similar values for these features: Philology (0.37±0.06) versus Natural Science (0.43±0.07). Higher metrics of the descriptive ratio of adjectives to nouns are also a characteristic of the texts of Natural Science (0.29±0.05), while Mathematical texts demonstrate a low descriptive ratio – 0.23±0.05. The metrics of Philological texts (0.25±0.04) in this respect are similar to the metrics of the Mathematical texts, rather than the ones of Natural Science ones (see Figure 4b).

Figure 4. a) Ratio of verbs to nouns; b) Ratio of adjectives to nouns
Рисунок 4. a) Отношение глаголов к существительным; b) Отношение прилагательных к существительным



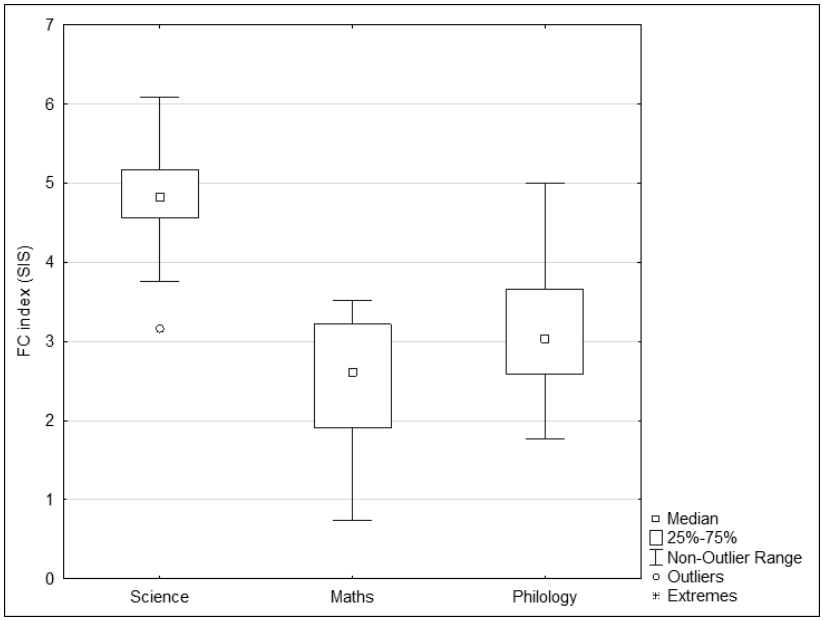
(a)



(b)

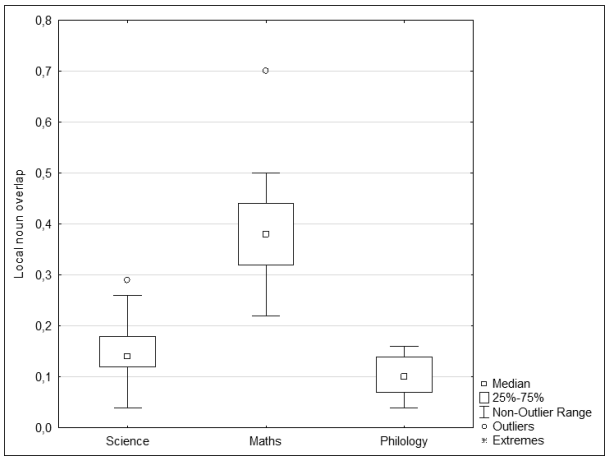
The Flesch-Kincaid index (SIS) is highest in natural science texts (4.83±0.59), while the metrics are quite similar in in Mathematical (2.51±0.81) and Philological (3.18±0.82) texts (see Figure 5).

Figure 5. Flesch-Kincaid (SIS)
Рисунок 5. Индекс Флеша-Кинкейда (SIS)

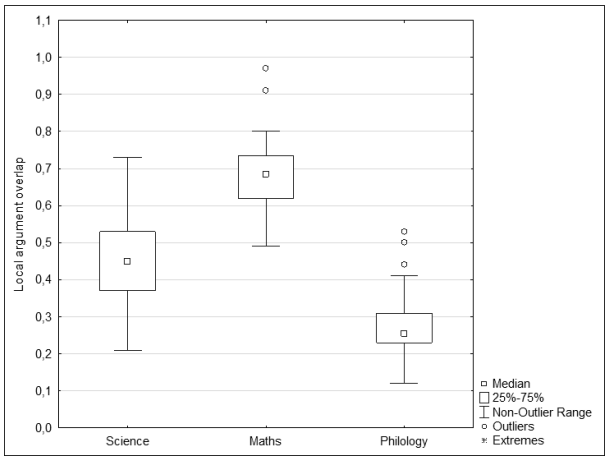


Mathematical texts demonstrate significantly higher metrics of local noun overlap (LNO=0.39±0.10) and local argument overlap (LAO=0.69±0.12). The values of these features are the lowest in the Philological texts (LNO = 0.10 ± 0.04, LAO = 0.28 ± 0.10), while the metrics for Natural Science texts occupy an intermediate position with LPS= 0.15 ± 0.06, and LAO = 0, 45±0.13) (see Figures 6a and 6b).

Figure 6. a) Local noun overlap; b) Local argument overlap
Рисунок 6. a) Локальный повтор существительных; b) Локальный повтор аргумента



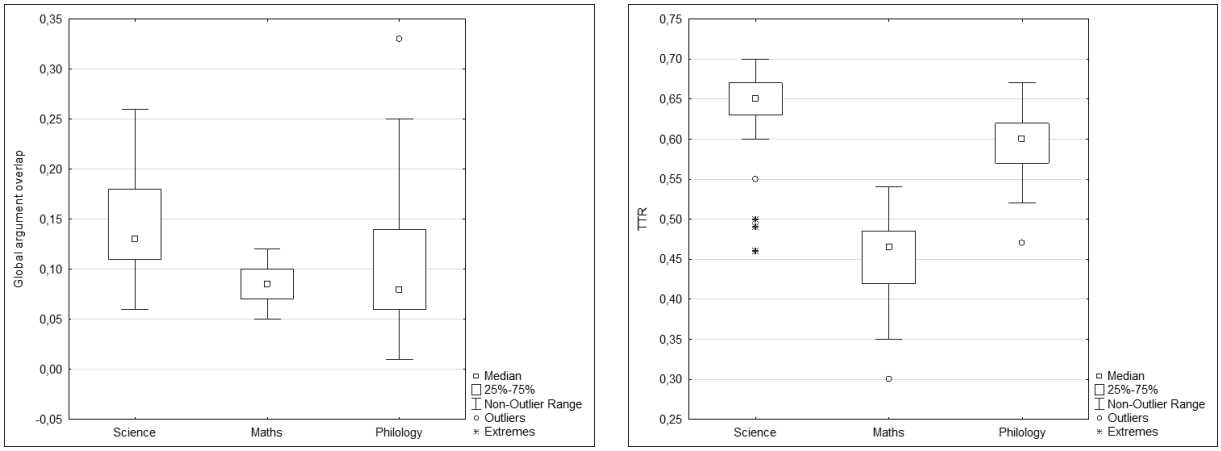
(a)



(b)

Figures 7 a and b show differences in metrics of global argument overlap (0.14±0.05) and lexical diversity (TTR) (0.64±0 .05). Global noun overlap in the Mathematical texts (0.08±0.02) are slightly lower than in the Philological texts (0.11±0.07), while the difference in the values of TTR of the Philological (0.60±0.04) and Mathematical (0.45±0.06) texts is higher.

Figure 7. a) Global argument overlap; b) TTR
Рисунок 7. a) Глобальный повтор аргумента; b) Лексическое разнообразие



(a) (b)

The identified linguistic features of educational texts were used to develop a mathematical model for profiling texts of the three discourses. To design a predictive model, we employed a discriminant analysis, one of the most validated multivariate methods in style studies (Andreev, 2010: 100–110). Discriminant analysis was also used in attribution studies (i.e., authorship identification; Baayen, Halteren & Tweedie, 1996: 121–132; Holmes, Forsyth, 1995: 111–127; Stamatatos, Fakotakis & Kokkinakis, 2001: 193–214).

For the profiling technique, we used the 12 statistically significant features from Table 2. We considered a backward stepwise Discriminant Analysis that retained 5 variables (see Table 3). The discriminant analysis of the 77 texts used for training showed the following results: Wilks' Lambda $\lambda = .03821$, $F\ (10.140)=57.619$, $p < .001$. The values of λ Wilks close to 0 indicate good discrimination of the contrasted objects. Based on the values of λ and F-criterion, we confirm the accuracy of the classification.

Table 3. Discriminant Analysis Results
Таблица 3. Результаты дискриминантного анализа

Features		λ Wilks'	λ Partial	*F*	*p*
1	Mean sentence length	0.067	0..760	12.633	< .001
2	Mean verbs per sentence	0.108	0.474	44.368	< .001
3	Mean adjectives per sentence	0.068	0.749	13.399	< .001
4	Local argument overlap	0.078	0.657	20.838	< .001
5	Global argument overlap	0.079	0.643	22.233	< .001

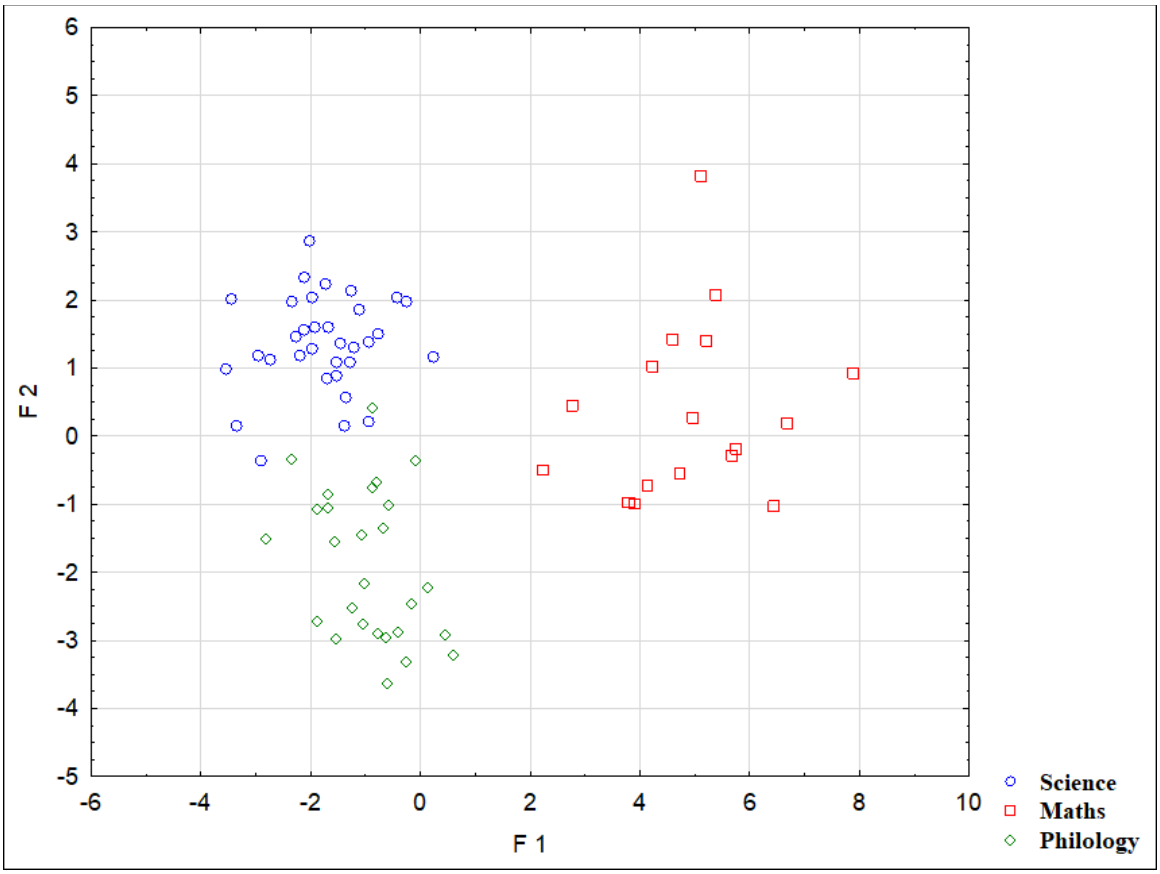
Given the test set of 10 texts, the accuracy of our model proved to be as high as 90% since 9 out of the 10 tested texts were correctly classified (see Table 4).

Table 4. Classification matrix
Таблица 4. Классификационная матрица

Sub-corpora	Projected Classifications			
	Accuracy (%)	Science	Maths	Philology
Science	100.0%	4	0	0
Maths	100.0%	0	3	0
Philology	66.7%	1	0	2
Total	90.0%	5	3	2

The scatterplot of the canonical values for canonical roots enables the identification of the contribution of each discriminant function in discriminating texts of each sub-corpora. As it can be seen from the diagram from Figure 8, the canonical function 1 (F1) differentiates Mathematics texts from philology and natural science texts: the higher the value of F1, the more likely it is that the text is Mathematical. Canonical function 2 (F2) enables us to differentiate Philology texts from Mathematics and Natural Science texts: the lower the value of F2, the more likely it is a Philological text. While inspecting the scatterplot, if the values of both canonical functions are negative, then the text is more likely to be classified as Philological; if the values of both functions are higher than zero, then the text is more likely to be classified as Mathematical.

Figure 8. Scatterplot of canonical values for canonical roots
Рисунок 8 Диаграмма рассеяния канонических значений для канонических корней



Based on the values of the standardized coefficients of canonical functions (see Table 5), we can define the impact of the linguistic features of the text on the values of canonical functions 1 and 2. Judging by the coefficients, the following linguistic features have the greatest influence on these functions: mean sentence length, mean verbs per sentence, and mean adjectives per sentence.

Table 5. Standardized coefficients of canonical functions
Таблица 5. Стандартизированные коэффициенты канонических функций

Text features	Acronym	Canonical Functions	
		F1	F2
Mean sentence length	M(Sen/L)	1.081	0.095
Mean verbs per sentence	M(VB/Sen)	-0.915	0.723
Mean adjectives per sentence	M(JJ/Sen)	-0.728	0.350
Local noun overlap	LocalNNOver	0.716	0.335
Global argument overlap	GlobalArgOver	-0.598	-0.035

Thus, the formulas for classifying texts by subject discourse are as follows:

F(Science) = -51.46+0.74*M(Sen/L)+47.67*M(VB/Sen)+21.75*M(JJ/Sen)+18.59*LocalNNOver+26.48* GlobalArgOver

F(Maths) = -51.21+7.65*M(Sen/L)+5.78*M(VB/Sen)+(-7.56)*M(JJ/Sen)+97.95*LocalNNOver+(-44.34)* GlobalArgOver

F(Philology) = -23.77+1.31*M(Sen/L)+28.55*M(VB/Sen)+11.93*M(JJ/Sen)+9.72*LocalNNOver+19.63* GlobalArgOver

Discussion

Text analysis showed that the educational texts from the three discourses exhibited statistically significant differences. For example, Natural Science texts differ from classroom texts on the Russian language and Mathematics by having longer sentences and higher nominative and descriptive ratios. The latter is probably caused by differences in their functions: Natural Science texts are supposed to create a holistic picture of the world and broaden the readers' horizons. An additional specific feature of Natural Science textbooks considers the constituency of the sample that includes texts from natural history, social science, and historical facts; as such, longer and more complex sentences are more frequently encountered.

On average, the number of verbs and nouns per sentence is higher in texts of the Natural Science textbooks. This argues that the authors of these texts draw attention to the subject or object of the action (higher frequency of content words), as well as to the internal structure of events (high ratio of verbs) (see Seifart, Danielsen, Meyer, Nordhoff et al., 2012: 10). Natural science texts contain more narrations of events and more descriptions of facts than opinions; in contrast, Philological texts contain more opinions.

Mathematics textbooks have a higher nominative ratio and a larger number of nouns per sentence compared to Philological texts (see Figure 3) and the mean word length in Mathematical textbooks is lower than in the contrasted discourses. Moreover, Mathematical texts have a low lexical diversity (see Figure 7b) since Mathematics operates with specific terms (traditionally denoted by nouns), and the use of synonyms in this type of discourse is either not recommended or impossible. This can also explain the higher values of local argument overlap in Mathematical texts when contrasted to texts from the two other discourses.

Educational texts in the textbooks used to teach the Russian language in elementary school differ from texts in other discourses by having shorter sentences, arguable since a large proportion of texts are intended for memorization and further reproduction. Short sentences are necessary to develop basic skills in writing and spelling skills.

Conclusion

The academic texts from three subject discourses (i.e., Philological, Mathematical, and Natural Sciences) exhibit statistically significant differences on 12 linguistic features, namely: sentence length, word length, mean nouns per sentence, mean verbs per sentence, mean adjectives per sentence, local noun overlap, local argument overlap, global argument overlap, nominative ratio, descriptive ratio, Flesch-Kincaid index (SIS), and lexical diversity (TTR). These differences are caused by the changes in the functions of these texts and are manifested in their lexical and syntactic levels.

Based on the discriminant analysis, we designed a model of text profiling that includes 5 linguistic features, namely: mean sentence length, mean verbs per sentence, mean adjectives per sentence, local noun overlap, and global argument overlap. The automation of these features and the provided formulas for classification enable the design and development of text profilers demanded for textbook writing and editing. Our model also contributes to the design of a quantitative linguistic typology of Russian academic texts.

Corpus Materials

Moro, M. I., Bantova, M. A., Bel'tyukova, G. V. et al. (2012). *Matematika. 2 klass. Ucheb. dlya Obshcheobrazovat. Organizacij v 2 ch. Ch. 1, 6-e izd.* [Mathematics. Grade 2. Textbook for secondary schools in 2 parts. Part 1, Edition 6.], Prosveshchenie, Moscow, Russia. ISBN: 978-5-09-028297-0 *(In Russian)*

Moro, M. I., Bantova, M. A., Bel'tyukova, G. V. et al. (2012). *Matematika. 2 klass Ucheb. dlya Obshcheobrazovat. Organizacij v 2 ch. Ch. 2, 6-e izd.* [Mathematics. Grade 2. Textbook for secondary schools in 2 parts. Part 2, Edition 6.], Prosveshchenie, Moscow, Russia. ISBN: 978-5-09-028297-0 *(In Russian)*

Peterson, L. G. (2017). *Matematika. 2 klass. V 3 ch. Chast' 1, Izd. 5-e* [Mathematics. Grade 2. Textbook in 3 parts. Part 1, Edition 5], Yuventa, Moscow, Russia. ISBN: 978-5-9963-3238-0 (In *Russian)*

Peterson, L. G. (2017). *Matematika. 2 klass. V 3 ch. Chast' 2, Izd. 5-e* [Mathematics. Grade 2. Textbook in 3 parts. Part 2, Edition 5], Yuventa, Moscow, Russia. ISBN: 978-5-9963-3239-7 (In Russian*)*

Peterson, L. G. (2017). *Matematika. 2 klass. V 3 ch. Chast' 3, Izd. 5-e* [Mathematics. Grade 2. Textbook in 3 parts. Part 3, Edition 5], Yuventa, Moscow, Russia. ISBN: 978-5-9963-3240-3 *(In Russian)*

Dmitrieva, N. Ya., Kazakov, A. N. (2021). *Okruzhayushchij mir: Uchebnik dlya 2 klassa v 2 ch. Ch. 1.* [Science. Grade 2. Textbook in 2 parts. Part 1.], Prosveshchenie, Moscow, Russia. ISBN: 978-5-0908-5490-0 *(In Russian)*

Dmitrieva, N. Ya., Kazakov, A. N. (2021). *Okruzhayushchij mir: Uchebnik dlya 2 klassa v 2 ch. Ch. 2.* [Science. Grade 2. Textbook in 2 parts. Part 2.], Prosveshchenie, Moscow, Russia. ISBN: 978-5-0908-5488-7 *(In Russian)*

Ivchenkova, G. G., Potapov, I. V. (2018). *Okruzhayushchij mir: Uchebnik dlya 2 klassa v 2 ch. Ch. 1.* [Science. Grade 2. Textbook in 2 parts. Part 1.], Astrel', Moscow, Russia. ISBN: 978-5-358-19400-7 *(In Russian)*

Ivchenkova, G. G., Potapov, I. V. (2018). *Okruzhayushchij mir: Uchebnik dlya 2 klassa v 2 ch. Ch. 2.* [Science. Grade 2. Textbook in 2 parts. Part 2.], Astrel', Moscow, Russia. ISBN: 978-5-358-19903-3 *(In Russian)*

Ramzaeva, T. G. (2022). *Russkij yazyk. 2 kl. v 2 ch. Ch. 1.* [Russian language. Grade 2. Textbook in 2 parts. Part 1], Drofa, Moscow, Russia. ISBN: 9785090793919 *(In Russian)*

Ramzaeva, T. G. (2022). *Russkij yazyk. 2 kl. v 2 ch. Ch. 2.* [Russian language. Grade 2. Textbook in 2 parts. Part 2.] Drofa, Moscow, Russia. ISBN: 978-5-09-087981-1 *(In Russian)*

Solovejchik, M. S., Kuz'menko, N. S. (2021). *Russkij yazyk. 2 klass.Uchebnik v 2 ch. Ch. 1.* [Russian language. Grade 2. Textbook in 2 parts. Part 1.], Prosveshchenie, Moscow, Russia. ISBN: 978-5-09-081119-4 *(In Russian)*

Solovejchik, M. S., Kuz'menko, N. S. (2021). *Russkij yazyk. 2 klass.Uchebnik v 2 ch.*

Ch. 2. [Russian language. Grade 2. Textbook in 2 parts. Part 2.], Prosveshchenie, Moscow, Russia. ISBN: 978-5-09-081121-7 *(In Russian)*

References

Andreev, V. S. (2010). Methods of quantitative style research in linguistics: a multidimensional approach, *Izvestiya Smolenskogo gosudarstvennogo universiteta*, 3 (11), 100–110. *(In Russian)*

Baayen, R. H., Halteren, H. and Tweedie, F. J. (1996). Outside the cave of shadows: using syntactic annotation to enhance authorship attribution, *Literary and Linguistic Computing,* 11 (3), 121–132. *(In English)*

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*, John Benjamins, Amsterdam. *(In English)*

Corlatescu, D., Ruseti, Ș. and Dascalu, M. (2022). ReaderBench: Multilevel analysis of Russian text characteristics, *Russian Journal of Linguistics,* 26, 2, 342–370, available at: URL: https://journals.rudn.ru/linguistics/article/view/31328 (Accessed 5 March 2022). https://doi.org/10.22363/2687-0088-30145 *(In English)*

Crossley, S. A., Varner, L. K., Roscoe, R. D. and McNamara, D. S. (2013). Using Automated Indices of Cohesion to Evaluate an Intelligent Tutoring System and an Automated Writing Evaluation System, *Artificial Intelligence in Education*, 269–278. https://doi.org/10.1007/978-3-642-39112-5_28 *(In English)*

Flesch, R. (1948). A new readability yardstick, *Journal of Applied Psychology,* 32 (3), 221–233. *(In English)*

Gatiyatullina, G., Solnyshkina, M., Solovyev, V., Danilov, A., Martynova, E. and Yarmakeev, I. (2020). Computing Russian Morphological distribution patterns using RusAC Online Server, *13th International Conference on Developments in eSystems Engineering (DeSE)*, 393–398, available at: https://ieeexplore.ieee.org/document/9450753.Coh-Metrix (Accessed 5 March 2022). http://doi.org/10.1109/DeSE51703.2020.9450753 *(In English)*

Graesser, A. C., McNamara, D. S., Louwerse, M. M. and Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language, *Behavior Research Methods, Instruments, & Computers*, 36 (2), 193–202. http://doi.org/10.3758/bf0319556 *(In English)*

Holmes, D. and Forsyth, R. (1995). The Federalist revisited: New directions in authorship attribution, *Literary and Linguistic Computing*, 10 (2), 111–127. *(In English)*

Kiselnikov, A. S. (2015). K probleme kharakteristik teksta: chitabel'nost', ponyatnost', slozhnost', trudnost' [To the problem of text characteristics: readability, clarity, complexity, difficulty], *Filologicheskie nauki. Voprosy teorii i praktiki,* 11 (53), 79–84. *(In Russian)*

Malvern, D., Richards, B., Chipere, N. and Durán, P. (2004). *Traditional Approaches to Measuring Lexical Diversity*, Palgrave Macmillan, London, UK. https://doi.org/10.1057/9780230511804 *(In English)*

Martynova, E., Solnyshkina, M. I., Merzlyakova, A. and Gizatulina, D. (2020). Leksicheskie parametry uchebnogo teksta (na materiale tekstov uchebnogo korpusa russkogo yazyka) [Lexical parameters of academic text (based on the texts of Academic corpus of the Russian language)], *Philology and culture,* 3 (61), 72–80, available at: http://www.philology-and-culture.kpfu.ru/?q=node/2728 (Accessed 5 March 2022). *(In Russian)*

McNamara, D. S., Graesser, A. C. and Louwerse, M. M. (2012). Sources of text difficulty: Across genres and grades, in Sabatini, J. P., Albro, E. and O'Reilly, T. (eds.), *Measuring up: Advances in how we assess reading ability*, 89–116. *(In English)*

McNamara, D., Graesser, A., McCarthy, P. and Cai, Z. (2014). *Automated Evaluation of Text and Discourse with Coh-Metrix*, Cambridge University Press, Cambridge, UK. http://doi.org/10.1017/CBO9780511894664 *(In English)*

Oborneva, I. V. (2006). Avtomatizirovannaya otsenka slozhnosti uchebnyh tekstov na osnove statisticheskih parameters [Automated assessment of the complexity of educational texts based on statistical parameters], Abstract of Ph.D. dissertation, Moscow City University, Moscow, Russia. *(In Russian)*

Seifart, F., Danielsen, S., Meyer, R., Nordhoff, S., Pakendorf, B., Witzlack-Makarevich, A. and Zakharko, T. (2012). *The relative frequencies of nouns, pronouns, and verbs*

cross-linguistically Applicant, available at: https://www.semanticscholar.org/paper/The-relative-frequencies-of-nouns-%2C-pronouns-%2C-and-Seifart-Danielsen/cd52cd7091fee4b1781c16a51fe58f87ba642c1c (Accessed 5 March 2022). *(In English)*

Sirotinina, O. B. (2009). *Spoken language within the system of functional styles of the Russian literary language: grammar*, Librekom, Moscow, Russia. *(In English)*

Solnyshkina, M. I. and Kiselnikov, A. S. (2015). Slozhnost' teksta kak funktsiya leksicheskih parametrov (na materiale uchebnyh tekstov na russkom yazyke [Text Complexity: Chronology of Russian applied linguistics studies], *Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya*, 6 (38). *(In Russian)*

Solnyshkina, M. I., Harkova, E. V. and Kazachkova, M. B. (2020). The Structure of Cross-Linguistic Differences: Meaning and Context of 'Readability' and its Russian Equivalent 'Chitabelnost', *Journal of Language and Education*, 6 (1), 103–119. https://doi.org/10.17323/jle.2020.7176 *(In English)*

Solnyshkina, M. I., Kazachkova, M. B. and Harkova, E. V. (2020). Cifrovye tekhnologii izmereniya slozhnosti tekstov kak instrument upravleniya kachestvom obucheniya chteniyu na anglijskom yazyke [Digital technologies for measuring text complexity as a tool for managing the quality of teaching reading in English], *Foreign languages at school*, 3, 15–21, available at: https://www.elibrary.ru/item.asp?id=42609743 (Accessed 5 March 2022). *(In Russian)*

Solnyshkina, M., McNamara, D. and Zamaletdinov, R. (2022). Natural language processing and discourse complexity studies, *Russian Journal of Linguistics*, 26 (2), 317–341. *(In English)*

Solnyshkina, M. I., Solovyev, V. D., Gafiyatova, E. V. and Martynova, E. V. (2022) Slozhnost' teksta kak mezhdisciplinarnaya problema [Text complexity as an interdisciplinary problem], *Issues of cognitive linguistics*, 1, 18-40. *(In Russian)*

Solovyev, V. D., Ivanov, V. V. and Akhtiamov, R. B. (2019). Dictionary of abstract and concrete words of the Russian language: A methodology for creation and application, *Journal of research in applied linguistics*, 10, 215–227. *(In English)*

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

Stamatatos, E., Fakotakis, N. and Kokkinakis, G. (2001). Computer-based authorship attribution without lexical measures, *Computers and the Humanities*, 35 (2), 193–214. *(In English)*

Vakhrusheva, A. Ya., Solnyshkina, M. I., Kupriyanov, R. V., Gafiyatova, E. V. and Klimagina, I. O. (2021). Lingvisticheskaya slozhnost' uchebnyh tekstov [Linguistic complexity of academic texts], *Voprosy zhurnalistiki, pedagogiki, yazykoznaniya: elektronnyj zhurnal*, 40 (1), 89–99, available at: http://jpl-journal.ru/index.php/journal/article/view/78 (Accessed 5 March 2022). *(In Russian)*

Zherebtsova, Zh. I. (2007). *Ispol'zovanie informatsionnoj struktury predlozheniya v obuchenii inostrannyh studentov-nefilologov chteniyu russkih uchebno-nauchnyh tekstov* [The use of information structure of the sentence in teaching foreign non-philological students to read Russian academic and research texts], Ph.D. Thesis, Herzen State University, St.Petersburg, Russia. *(In Russian)*

Zhuravlev, A. F. (1988). An experience of quantitative and typological investigation of spoken registers, *Varieties of urban spoken language: a collection of research articles. Raznovidnosti gorodskoy ustnoy rechi*, Nauka, Moscow, Russia, 84–150. *(In English)*

Конфликты интересов: у авторов нет конфликта интересов для декларации.

Conflicts of Interest: the authors have no conflict of interest to declare.

Roman V. Kupriyanov, Candidate of Psychology, Chief Researcher, Text Analytics Research Laboratory, Institute of Philology and Intercultural Communication, Kazan Federal University; Associate Professor, Department of Social Work, Pedagogy and Psychology, Kazan National Research Technological University, Kazan, Russia.

Роман Владимирович Куприянов, кандидат психологических наук, доцент старший научный сотрудник, НИЛ «Текстовая

аналитика», Институт филологии и межкультурной коммуникации, Казанский (Приволжский) федеральный университет; доцент кафедры социальной работы, педагогики и психологии, Казанский национальный исследовательский технологический университет (КНИТУ), Казань, Россия.

Marina I. Solnyshkina, Doctor of Philology, Head and Chief Researcher, Text Analytics Research Laboratory, Professor of the Department of Theory and Practice of Teaching Foreign Languages, Institute of Philology and Intercultural Communication, Kazan Federal University, Kazan, Russia.
Марина Ивановна Солнышкина, доктор филологических наук, профессор, профессор кафедры теории и практики преподавания иностранных языков, руководитель и главный научный сотрудник, НИЛ «Текстовая аналитика», Институт филологии и межкультурной коммуникации, Казанский (Приволжский) федеральный университет, Казань, Россия.

Mihai Dascalu, Ph.D. (CS), Ph.D. (Edu), Professor, Dr., Department of Computers, Polytechnic University of Bucharest, Bucharest, Romania.
Михай Даскалу, доктор наук (Информационные технологии, Образование), профессор, профессор кафедры вычислительной техники, Бухарестский политехнический университет, Бухарест, Румыния.

Tatyana A. Soldatkina, Candidate of Philology, Chief Researcher, Research Laboratory “Expert Systems for Processing Language Structures and Vibroacoustics”, Kazan Federal University; Associate Professor of the Department of English Philology, Mari State University, Yoshkar-Ola, Russia.
Татьяна Альбертовна Солдаткина, кандидат филологических наук, доцент, старший научный сотрудник, НИЛ «Экспертные системы обработки языковых конструкций и виброакустика», Казанский (Приволжский) федеральный университет, доцент кафедры английской филологии, Марийский государственный университет, Йошкар-Ола, Россия.